

DISSERTATION

THE DIGITAL PRESERVATION OF RESEARCH AT COLORADO STATE UNIVERSITY:
A CASE STUDY OF THREE DEPARTMENTS

Submitted by

Edgar U. Peyronnin

Department of Journalism and Media Communication

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2015

Doctoral Committee:

Advisor: Peter B. Seel

Jeni Cross

Kris Kodrich

Dawn Paschal

Craig Trumbo

Copyright by Edgar Ursin Peyronnin 2015

All Rights Reserved

ABSTRACT

THE DIGITAL PRESERVATION OF RESEARCH AT COLORADO STATE UNIVERSITY: A CASE STUDY OF THREE DEPARTMENTS

Research workflows in higher education have converged onto digital formats. While the technology to store data has improved at an increasing pace, personal and organizational behaviors have not adapted as rapidly. The study used Diffusion of Innovation theory concepts within an Activity Theory construct and the Open Archive Information System to create a model for studying key areas of transformation. The model provides a new way to understand the complex set of issues that can inhibit data preservation. The key areas were determined by analysis of interviews, surveys and institutional data. The study used descriptive statistics and social network analysis to elaborate ways to transmit new data preservation attitudes and behaviors more effectively. The study proposes three temporal contexts – short-term, long-term and trans-generational. The data management plan requirement for National Science Foundation grant submissions was a powerful motivator. The study found opportunities for the institution to create group activities, such as workshops, that specifically include faculty with NSF grants and those who share other grant submission experience with them. The study also found that information technology staffs need to understand research problems from the researcher perspective better to overcome some trust issues. Finally, campus leadership needs to identify their role in addressing the issue for the long-term benefit of the institution. Strategic goals are an important first step. Building a robust digital preservation environment is an iterative process dependent on many perspectives. The goal of this research is to speed the process by developing a systems-level model for exposing problem areas.

ACKNOWLEDGEMENTS

I would like to thank my committee for their guidance and efforts through this personal journey. Each brought knowledge and perspective from diverse backgrounds to help me focus my ideas and develop my methodology. Each contributed in ways for which I am very thankful. I would particularly like to thank my advisor through the entire project, Dr. Peter B. Seel. Although the original research ideas are still the central focus of this paper, the final product is far different than what was initially envisioned. A good advisor points you toward your goal, a great advisor allows you take the scenic route. You were great, Pete. I would also like to thank faculty, friends and colleagues who encouraged me through all the long days. Earning a PhD is a huge endeavor, even without a full time job. Kind words or a well-timed quip can change the trajectory of any day. If you made me laugh, you helped me finish. To Zach, who listened to my ideas while we cycled over many mountain passes in Colorado. To my parents, Joe and Dorothy, who stressed a devotion to family and the importance of education. I know that they would be proud. To my sister, Dr. Ann Trent, who waited almost as long as I did to earn her PhD and showed me that it could be accomplished late in life. Thank you for the inspiration. To my brother, Joe Peyronnin, who, among other wisdom he has provided, gave me The Shortness of Life by Seneca. Life is never short if invested well. To my children Ben and Melissa, their spouses, Elizabeth and Craig, and my granddaughter, Landry, who give me optimism for the future and the knowledge that some of life's accomplishments can't really be measured, but certainly can be enjoyed. Finally to my partner for over 37 years, Barbara – I think we're done with school now. It couldn't have been accomplished without your love and support. You were always positive and inspiring. Thank you for saying "I do" so many years ago.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGEMENTS	iii
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER 1 INTRODUCTION	1
Preserving Knowledge During A Digital Dark Age	1
Evolution of Sustainable Digital Behavior And Responsibility	4
CHAPTER 2 BACKGROUND	7
Context and Relevant Research	7
Passing The Human Experience Forward.....	7
What Drives Innovation in the Rapidly Developing Digital World	16
The Open Archive Information System (OAIS).....	21
Library of Congress' Seven Digital Sustainability Factors	23
The National Institute of Health Data Sharing Policy	25
The National Science Foundation Data Plan	26
Council on Library and Information Resources pub154.....	29
Changing Researcher Workflows for the Digital Age.....	31
Theoretical models.....	33
Socializing The Digital World.....	33
Diffusion In The Rapidly Changing Digital World	35
The Act Of Making And Socializing Our Tools.....	41
Model and Research questions.....	45
CHAPTER 3 METHODS	47
Synopsis	47
Connected Research - Institutional Records	51
Initial Interviews – Create Survey Categories	51
Face-to-Face Surveys – Detailed Fact Finding.....	52
Final Interviews –Issues Requiring More Elaboration	53
CHAPTER 4 RESULTS AND ANALYSIS	54
Subject.....	54
Instrument	58

Division of Labor.....	60
Community	65
Rules	70
CHAPTER 5 DISCUSSION.....	81
Contextualizing Data Preservation	81
Subject.....	82
Instrument	82
Division of Labor.....	83
Community	84
Rules	84
Finding Answers	88
Recommendations.....	90
CHAPTER 6 CONCLUSION.....	96
Limitations	98
Future research.....	98
BIBLIOGRAPHY.....	101
APPENDICES	108
APPENDIX A - Initial Interview questions (Spencer, p. 26).....	109
APPENDIX B - Face to Face Survey Questions	111
APPENDIX C - Final Interview Questions	118
APPENDIX D - National Science Foundation Organization Chart	119
APPENDIX E - National Science Foundation Data Management Plan requirement excerpt....	121
APPENDIX F - Open Data Policy - Office of the President of the United States	124
APPENDIX G - Increasing Access to the Results of Federally Funded Scientific Research	136

LIST OF TABLES

Table 1 Library of Congress' Seven Digital Sustainability Factors.....	24
Table 2 National Science Foundation's Guidelines for Inclusion in a Data Management Plan...	27
Table 3 Number Of Tenure Track Faculty By Department (Institutional Research - Colorado State University, 2015).	55
Table 4 Award Dollars Received From External Sources By Department x 1,000 (Colorado State University Vice President for Research, 2015).....	55
Table 5 Student Enrollment By Department (Institutional Research - Colorado State University, 2015).	55
Table 6 Survey Response Rate	58
Table 7 Average Data State By Department.....	59
Table 8 File Management: Knowledge.....	59
Table 9 Good File Naming Habits	60
Table 10 Data Responsibility Average Response By Department	61
Table 11 Manual Or Automatic Backups	65
Table 12 Data Preservation Requirements.....	71
Table 13 Data Repository Locations.....	72
Table 14 Meta-Data Repository Location.....	72
Table 15 Data Management Plan And Researcher Action	73
Table 16 Weighted Out-Degree: All Federal Research Grants 1987 - 2014 By Study Group.....	75
Table 17 Weighted Out-Degree NSF Research Grants 1987 - 2014 By Study Group.....	76
Table 18 Comparative NSF Influence On Data Deposit	79
Table 19 Sample Incidence Matrix	85
Table 20 Adjacency Matrix: Digital Preservation Attributes By Digital Preservation Attributes	86
Table 21 Adjacency Matrix: Faculty By Faculty.....	86

LIST OF FIGURES

Figure 1 Life expectancy of media through the ages (Conway, 1996).....	12
Figure 2 OAIS Model(Council of the Consultative Committee for Space Data Systems, 2011, pp. 4-1).....	22
Figure 3 Engeström’s representation of Vygotsky's Activity (Engeström, Miettinen, & Punamäki-Gitai, 1999, p. 30).....	42
Figure 4 Engeström enhancement of Vygotsky's AT representation (Engeström et al., p. 31)....	43
Figure 5 Proposed transformation of research data to a sustainable information package within Activity Theory construct.	45
Figure 6 Sociometric Star.	48
Figure 7 Balanced and unbalanced structures with multiple connections between “B” and “C”. 49	
Figure 8 Organizational Trust Inventory – Anthro	62
Figure 9 Organizational Trust Inventory – Atmos.....	62
Figure 10 Organizational Trust Inventory – CIS	63
Figure 11 The most time a data incident in which critical research/data was lost cost	64
Figure 12 Backup Frequency	65
Figure 13 What data do researchers expect to share from their research? (During research)	69
Figure 14 What data do researchers expect to share from their research? (After research)	70
Figure 15 Weighted Out-Degree for Federal Grants 1987 - 2014 - Target Group	74
Figure 16 Weighted Out-degree for NSF University Grants 1987 - 2014 - Study Group.....	75
Figure 17 Eigenvector Centrality Distribution – Federal grants 1987 - 2014	77
Figure 18 Successful Grant Submissions for three Departments since 2011	78
Figure 19 Indirect connections to NSF	80
Figure 20 Study Model Using Activity Theory	81

RESEARCHER'S PERSPECTIVE

As an information technology professional in higher education for over two decades, I have supported research in our college exploring a wide-range of investigations, from agricultural marketing questionnaires to DNA sequencing food crops. Research funding sources include industry groups, US Department of Agriculture, National Institute of Health, and National Science Foundation. I am often confronted with advising my faculty on how to deal with the data they generate at the beginning, during, or at the end of their careers. While the advice I give is technically correct, researchers struggle to habituate for a myriad of reasons analogous to adopting healthy habits in life. “It’s too troublesome, I don’t have time, I am getting conflicting advice”. Mediated communication is an integral component of any solution. I have built data centers with identical, geographically separated, and mirrored data stores. I have chaired a university committee to find a solution for centralizing data centers. The committee created a business model, campus policy and technology for a campus cloud data center which went into production in 2012 on our campus. I have participated with several other ad-hoc committees from our libraries on our campus to help define the problem and write National Science Foundation grants to fund our campus digital repository. I have also investigated commercial cloud systems. Throughout it all, I recognized the need for a fundamental change in personal habits, institutional policy, and business processes to complement the effort by technologists who build the virtual environments our knowledge occupies today. These solutions must be designed so that they can adopt and evolve as the technology progresses. This study is an attempt to answer very important questions for the digital era.

CHAPTER 1 INTRODUCTION

Experts have understood the risks associated with sustaining a digital world for several years. The goal of this study is to find ways that digital preservation activity can be communicated to researchers to improve the permanency of their research data. The research focused on activities in an institutional environment that result in the creation of sustainable information packages (SIP) for research data. SIPs are the input to the Open Archive Information System (OAIS) model. The OAIS is an International Standards Organization (ISO) standard that constitutes the framework for the new digital environment. The research used a case study method to answer the question “How can organizational resources be effectively communicated to researchers to improve their file management skills?” The study conducted interviews, face-to-face surveys, and used both institutional and government data for the analysis. The research developed concepts from the transcriptions, used descriptive statistics, and social network analysis. Finally, the study offered insights and recommendations to help foster long-term preservation of institutional research that may be useful to other institutions.

Preserving Knowledge During A Digital Dark Age

Today, mediated communications and knowledge have converged onto digital formats that have been adopted by almost every society on the planet. Although the implications of the digital transformation have been studied from many perspectives, its impact may not be fully understood for generations to come. Since Turing’s (1937) seminal paper “On Computable Numbers, with an Application to the Entscheidungs problem” proposed a new digital model over seventy-five years ago, the accelerating rate of digital data growth has increased both the number of extraordinary opportunities for discovery and the risk of unintentionally removing priceless

information sources from the archive of collective human knowledge with a keystroke. While the technology to store data (hardware and software) has improved at an increasing pace, personal and organizational behaviors have been much slower to adapt to the new environment. Data storage has become easier as the price of storage devices has plummeted while data preservation has become harder because of the task's magnitude and unseen, abstract nature.

Data storage and data preservation are two very different concepts. Storage is dependent on a "place" such as a disk drive while preservation is an "act" dependent on human behavior. Storage relies heavily on technology that allows users to store more data in more ways on increasingly smaller devices. As devices and storage media evolve, data are left on outdated media or file formats – bits locked in a bottle afloat at sea. This "digital amnesia" is certainly avoidable, but only if the data are properly attended. Threats to data preservation, such as media obsolescence, can be resolved through automation, but ultimately depends on human intervention to assess, describe, and prioritize digital artifacts. Preservation implies "stewardship" to maintain its readability, content, and meaning. Digitally stored information is an intangible asset whose cost to accumulate is inexpensive and becoming more so as technology progresses. In contrast, digitally preserved information is more complex, difficult and becoming more expensive to sustain. According to Hedstrom (1997), digital preservation includes "the planning, resource allocation, and application of preservation methods and technologies to ensure that digital information of continuing value remains accessible and usable" (p. 190). Thus, data preservation requires that data be described using metadata based on standards and migrated regularly in accordance with a plan to accommodate hardware and software upgrades.

The costs associated with preserving digital artifacts have been easy to ignore or pass on during the early years of the digital transformation. However, it reached a tipping point at

approximately the beginning of the millennium with the broader realization that society was condemning itself to a “digital dark age”(Kuny, 1997, p. 1). What role does communication have in developing the ethical constructs, spreading the understanding of responsibility, and mediating solutions for promoting positive digital health? The National Science Foundation implemented a data management plan requirement on January 18, 2011, in order to increase innovation by opening access to research data as soon as possible (National Science Foundation, 2013). The NIH required a similar standard in 2003 and the White House Office of Science and Technology Policy issued a memorandum directing “each Federal Agency with over \$100 million in annual conduct of research and expenditures to develop a plan to support increased public access” (Office of Science and Technology Policy, 2013). As standard digital practices evolve, they also need to adapt to rapidly evolving technologies. The interchange between individuals, organizations, and the technology itself is critical to the evolution.

Research has shown that within-discipline influences are strong relative to institutional pressures for faculty. Professional relationships are extremely influential to faculty; in many ways, more than with the institution they work for. Before someone is willing to share research, there needs to be trust and common language so that the work is used in the context that it was intended. Academic fields and professional staff have nuanced cultures and jargon that can inhibit partnerships. Many universities provide resource (money, facility, and staff) to begin research, but require faculty to seek research funding and gifts to build their research program. Professional reputation is built by journal or book publishing and conference presentations which are largely discipline-centric. Recent NSF and NIH data management plan requirements affect some domains (e.g. sciences) more than others (e.g. arts). How are digitally sustainable practices communicated between disciplines and across an organization? Digital preservation standards

are in a formative stage. Digitally sustained practices require formalized trust relationships that information will be preserved. The interaction and mutual evolution with tools is part of a shared cognitive experience. The urgency to quickly find solutions increases the risk of settling on the wrong solution. Many of the answers are still neither fully formed nor articulated. The pace of change does not leave much time to negotiate the transactions. Communications and collaboration focused on the issue are essential to minimize these mistakes and to preserve information for future generations.

Evolution of Sustainable Digital Behavior And Responsibility

In an interview prior to this study, a biology professor at Colorado State University discussed research he conducted in the very early 1990's that had some inconsistent results and that were not useful for his project. After filing his report with the granting agency, he dutifully placed the logs with all the data on a shelf in his lab. In a separate project fifteen years later, he realized that some of this earlier research was actually valid based on new findings. He used the earlier data to strengthen findings in his new research. Would researchers' digital behavior and habits today permit the same success fifteen years from now?

Until the end of the twentieth century, researchers relied upon a process based on paper-trails of log books, charts, manuals and journals. Many scientists maintained their data for their entire lives in their labs, sometimes referring back to previous unsuccessful experiments that informed research years later. In 1991, Tim Berners-Lee and Robert Cailliau from CERN developed a means to use hypertext, the TCP/IP protocol, and Domain Naming Service to create a collaborative space (Berners-Lee & Fischetti, 1999). "The idea was to connect hypertext with the Internet and personal computers, thereby having a single information network to help CERN physicists share all the computer-stored information at the laboratory" (CERN, 2012). Since the

creation of the World Wide Web twenty-five years ago, university research has evolved from an activity with well-accepted practices and procedures for written documentation to a new, digitally based activity.

In 2001, a University of Southern California neurobiologist, Dr. Joe Miller, discovered he couldn't read magnetic tapes from the 1976 Viking landings on Mars. With the data in an unknown format, he had to track down printouts and hire students to retype everything. "All the programmers had died or left NASA," Miller said. "It was hopeless to try to go back to the original tapes" (Jesdanun, 2003, p. 10). In his re-examination of the remaining data, Dr. Miller discovered that there possibly were signs of life on Mars that earlier research had not understood. The gas emissions, previously thought to be simply the chemistry, were possibly a part of a circadian rhythm that could be attributable to life in the soil. Unfortunately, his team only had fragments of the record and corroborating proof may have been lost with the tapes. Alternatively, University of Sunderland's Dr. Dennis Wheeler used ships' logs dating back to the 1760's for data to contribute to his research into global warming discussion (Sunderland, 2009). Thus, handwritten data from 250 years ago can be read, but digital data can't be read after only twenty-five years.

For generations, at the end of their careers, researchers would routinely pass their collected works on to colleagues who shared their research interests. This distribution of lab books, reports, papers, and raw data could sit on these colleagues' shelves for years without any preservation activity as they incorporated it into their own research. Since the problem is so new and practices haven't been established, it is unlikely that faculty researchers organize and determine the value of their digital files in accordance with any sustainable guidelines today.

The digital age causes a re-evaluation of research workflow. Information is unreadable

without the correct hardware and software and can be trapped on media that cannot be accessed. For example, the value of any data stored on Apple 5E 5 ¼ inch floppies needs to be relatively high to someone in order to undertake the effort of migrating it to an updated format. It is extremely difficult to establish the value without some documentation of what is actually on the media. In many cases, only the researcher knows what is on the disks. Even if the means to read the disks is found, there is documentation, and the state of the disk is good, file organization practices may further inhibit a recovery project. Some researchers are meticulous and some thrive in the chaos of discovery. Kuny observed that there is a

demographic bulge of electronic materials coming into libraries and archives as the Baby Boom generation of authors and academics begin to wind down their careers and begin off-loading their materials to various libraries and archives (Kuny, 1997, p. 9).

This bulge will probably include quite a bit of problem data as described.

Standard archival activities, practices, policies, and training plans are instrumental for a successful digital sustainability program. Communication plays a vital role developing the materials, and campaign strategies for mediated channels to the public. The solution requires education and an understanding of both technology and human behavior. Who does the information benefit and whose responsibility is it to consciously ensure that the important parts are maintained for future generations? Berger maintains that while librarians are stewards, the originator is responsible for selection and prioritization (Berger, 2009). If this is the case, it is critical to understand researchers' digital preservation habits. The proper maintenance of society's digital knowledge base is increasingly more dependent on individual behaviors and attitudes.

CHAPTER 2 BACKGROUND

Context and Relevant Research

Passing The Human Experience Forward

Since the beginning of civilization, humans have passed knowledge on to succeeding generations to sustain a society's way of life. "Language in its fundamental forms is the symbolic expression of human intuition" (Sapir, 1921, p. 124). Children learned their family history from their parents and grandparents. Parents taught their children to build homes, plant seeds and hunt for food. The knowledge was passed from one generation to the next through oral traditions. Although the method was effective in that civilization survived, the information passed from one generation to the next was restricted to the human mind's capacity to remember. As a consequence, it limited the complexity of physical (buildings, and roads) and abstract (geometry, psychology, etc.) structures to the details of a concept that could be committed to memory. The oral tradition was insufficient to support intricate physical or conceptual structures. As the oral tradition transitioned to a written one, information could be stored, referred to, and passed from one generation to the next more easily. The expression of language in written form evolved in different ways as cultures adopted and adapted to their languages.

It is quite an illusion to imagine that one adjusts to reality essentially without the use of language and that language is merely an incidental means of solving specific problems of communication or reflection. The fact of the matter is that the 'real world' is to a large extent unconsciously built up on the language habits of the group. No two languages are ever sufficiently similar to be considered as representing the same social reality. The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached (Sapir & Mandelbaum, 1949, p. 45).

Logan states the way written language evolved from primitive morphology primarily into alphabetic/phonetic in the West (Logan, 2004) required different abstraction and analytic

cognitive skills creating the environment conducive to the development of objective logic and deductive reasoning. Like a digital system, the alphabet relies on encoding and decoding, combining characters to form phonemes and phonemes to form words and concepts. Logan then claims that science practiced in ancient China is different in part because of their alphabet's effect. One favorable characteristic of the Chinese logographic system is after thousands of years, it can be read by average people. It is extremely difficult for the modern reader to understand Beowulf.

Three thousand years ago, near the Babylonian city of Uruk, citizens learned how to inscribe clay to designate a trade of their property. This spread to the Egyptians with whom they traded, to the Phoenicians, and eventually to the Greek culture. Each step of the way, the written language was improved, from pictographs to logographs and finally to a phonetic alphabet in Western Europe and Middle East. The knowledge of humanity grew. Socrates' teachings in the oral tradition may have been forgotten had it not been for his student, Plato, capturing them in his dialogues. Plato's Phaedrus captures Socrates warnings that the written word would condemn humankind to a world without memory.

The specific which you have discovered is an aid not to memory, but to reminiscence, and you give your disciples not truth, but only the semblance of truth; they will be hearers of many things and will have learned nothing; they will appear to be omniscient and will generally know nothing; they will be tiresome company, having the show of wisdom without the reality.(Plato & Jowett, 1931, p. 402)

These warnings foreshadow similar warnings today about Google's effect on memory (Carr, 2011). Despite Socrates warnings, written words were resilient. Text could even be lost awaiting rediscovery many generations later. Many of the discoveries of the European Renaissance were simply "rediscoveries" of things known, rewritten, and preserved from earlier Greek and Roman civilizations largely by Catholic church scribes. As the writing process

improved so did communication. Humanity's depth and breadth of knowledge grew. The written tradition as an almost exclusive medium of choice for communication has lasted several thousand years until the beginning of the twenty-first century.

Innis (1972) describes the impact of the form of communication on a culture, society, or political organization. He concluded that the mediation of communication emphasizes either space or time by nations (empires). Those that emphasize space more efficiently communicate and control large numbers of their population. The continuity and stability of communications that emphasize time, limit an empire's ability to communicate in the present tense, but ideas are preserved for future generations.

The concepts of time and space reflect the significance of media to civilization. Media that emphasize time are those durable in character such as parchment, clay and stone...Media that emphasize space are apt to be less durable and light in character such as papyrus and paper (Innis, 1972, p. 7).

In the end, he concludes that the tension between the two will continue with a bias toward space. "The ability to develop a system of government in which the bias of communication can be checked and an appraisal of the significance of space and time can be reached remains a problem of empire and of the Western work"(Innis, 1972, p. 170). The bias in mediated communication continued to shift toward "space". As written communications evolved from clay and stone to papyrus and paper, an empire's boundaries were more efficiently extended. The need to administer and control larger territories more effectively through better communications technology continues to support "space". Communication in the digital age is instantaneous. However, even Innis' considerations of space and time may be seen in a different light as societies engage and shape the tools of the digital age to its needs. Seel describes the irony that, through archival efforts, once transitory space-bound media are now both archived

and discoverable. “These space-bound media are now actually less transitory through online search engines and digital archives” (Seel, 2012, p. 117).

The transition to digital incurs a fundamental change to the system. Dimitrova and Bugeja cited three criteria for archives that haven’t changed since the 15th century until the evolution of digital format: place, implement, and material. “All these factors have changed with the advent of the digital library, which exists in cyberspace and houses records owned by others that were created on software licensed by vendors and stored on files on servers not in the library” (Dimitrova & Bugeja, 2007, p. 2). In the future, libraries will no longer own the “documents”. Physical document repositories are disappearing across the country with increasing frequency. Instead, documents are held on file servers, sometimes proprietary, that conveniently deliver information to researchers. If a publishing company were to go out of business ten or twenty years ago, their journals would still be in readable format today in many libraries. If the same company went out of business today, questions about journal accessibility would and should be raised. The fire at the Library of Alexandria in ancient Egypt is still regarded to this day as a tragic event to society (Heller-Roazen, 2002, p. 133). Society confronts the possibility of a recurrence of this tragedy in our continuing headlong drive into the digital age.

In an increasingly digitized world, experts have just begun to develop the tools and structures to cope with our digitized daily lives. Policies and attitudes to sustain our digital infrastructure have lagged behind instruments that create it. How much data can there be? “In 2007 the amount of information created will surpass, for the first time, the storage capacity available” (IDC, 2007, p. 2). The International Data Corporation estimated that, in 2006, there were 161 exabytes (megabyte, gigabyte, terabyte, petabyte, exabyte) in digital data, or about three million times the information in books ever written (IDC, 2007). By 2013, IDC estimated

that there were 4.4 zettabytes (petabyte, exabyte, zettabyte) and that it would grow to 44 zettabytes by 2020. (EMC2, 2014, p. 4) The growth is driven by the evolution of connectable devices, known as the Internet of Things, and emerging markets. Approximately two-thirds of the data are created by consumers. The bits (binary digits) of a file can exist long after the means to open and read its contents are available. What of a future in which mountains of data are stored in meaningless zeroes and ones? Kuny (1997) described a “digital dark age” in which our documents, unlike those one thousand years ago, become obsolete every five to ten years and are hopelessly lost to future generations. The technological euphoria of the 1990’s has led to a digital flood in global societies lives today.

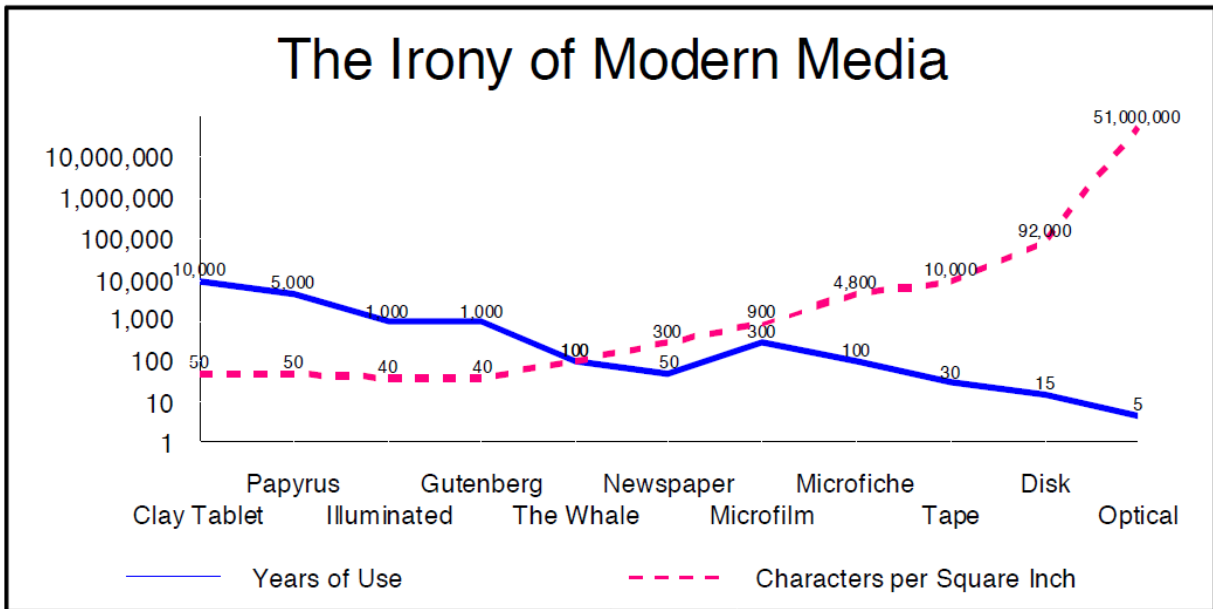
We seem at times, to be living in what Umberto Eco has called an ‘epoch of forgetting.’ Within this hyperbolic environment of technology euphoria, there is a constant, albeit weaker, call among information professionals for a more sustained thinking about the impacts of the new technologies on society (Kuny, 1997, p. 1).

Stille addressed concerns in the transition to digital and its impact on culture and society.

One of the great ironies of the information age is that, while the late twentieth century will undoubtedly have recorded more data than any other period in history, it will also almost certainly have lost more information than any previous era. A study done in 1996 by the Archives concluded that, at current levels, it would take approximately 120 years to transfer the backlog of non-text material (photographs, videos, film, audiotapes, and microfilm) to a more stable format (Stille, 2006, p. 5).

These records will be digitized once they are determined to be of sufficient value to someone. This person could be a professional archivist or someone with very little background. Individuals creating the digital record bear the greatest responsibility of ensuring that the artifact will last in the harsh digital environment. Institutional archives will need to commit resources to digital collections. Resources for preserving items not digitized will diminish as they are allocated to digital archives.

Once the transition from paper to digital format is complete, the media itself must be maintained. Conway (1996) analyzed the progression of media forms from the clay tablet to the CD and discussed the accompanying dilemma and irony.



Paul Conway, *Preservation in the Digital World*, 1996

Figure 1 Life expectancy of media through the ages (Conway, 1996).

...the capacity to record and store gives rise to one of the central dilemmas of recorded history: Our capacity to record information has increased exponentially over time while the longevity of the media used to store the information has decreased equivalently... The newest recording medium--optical disk--may indeed have a longer life than the digital recording surfaces that have gone before. It is likely, however, that today's optical storage media may long outlast the life of the computer system that created the information in the first place (Conway, 1996, p. 10).

Twenty years after his article, the CD and DVD are on the “digital-media” endangered list. People now use flash drives or cloud storage. Although cloud storage may be more reliable, an element of personal control is lost in the bargain. Belying its stable appearance to the average user, cloud storage itself is extremely dynamic. It runs on thousands of spinning disks in racks running on redundant power. Drives fail regularly but the system is protected by redundant

disks. The power that supplies the system is redundant, but that does not mean it is beyond failure due to natural or human caused disaster. Our cyberinfrastructure is a lucrative target and relatively open to threat (Council on Foreign Relations, 2014).

Berman (2008) projected four significant cyber infrastructure trends: more digital data are being created than there is storage to host it; more policies and regulations require the access, stewardship, and/or preservation of digital data; storage costs for digital data are decreasing; increasing commercialization of digital data storage and services. Third party commercialized digital storage relinquishes some control and lacks the fixity of documents written on paper. It conforms to Innis's concept of space over time. It is very difficult to change a written document without detection. It is relatively "fixed" in place. Electronic documents can be altered relatively easily leaving no trace of the original content. While a cynical observer may dismiss conspiracies, in the age of intentional sound manipulation, and photo editing, the risk is much greater than in the past. Even if malfeasance is not intended, content can be lost with each software upgrade. Content creators confront increasingly complex decisions to ensure the intended meaning of their words, sounds, images, and ideas remain for future generations.

Converged media is a phenomenon with greater significance in a digital paradigm than written or oral. The 2008 election of Barack Obama was an historic event. His staff, known for their use of new media, included messaging in video games to attract the 18 – 34 year old voters (Montagne, 2008). How shall these types of messages be preserved? Ad-hoc actions preserve some of these, but future generations may have to rely on the written, third-person accounts about the campaign. How is content and meaning from e-media that contains text, moving and still images, sound, web links, and other embedded code whose ownership and rights cannot be disambiguated preserved?

Digital sustainability helps define how individuals, organizations or societies, prevent loss of digital knowledge due to the rapid pace of change in the hardware and software upon which it depends. The concept requires users to consider knowledge lost as a consequence of the pace of change within the rapidly growing digital structures many institutions are creating. Information stored physically on obsolete devices such as floppy, zip or jazz drives, CD's, as well as created with obsolete software are only a part of the loss. Digital sustainability requires active engagement by all who use technology in their lives. The loss is not exclusively the result of computer and application obsolescence. In a converged digital world, it includes the "mashed-up" information that permeates the world today. Who owns it and how should it be organized? Each person contributes to the problem daily. Who is responsible or accountable to understand the issue? Who gets to write the history? The digital repositories society is building are the foundation for our future generations' knowledge and historical record.

Librarians and archivists have led research in digital sustainability. However, as Berger (2009) clearly points out from an ethical perspective, while professionals in libraries and archives bear responsibility for stewardship, it is the creator of each artifact who should have the first opportunity to establish the value of each object which helps ensure that it is deposited into an archive and maintained in perpetuity. Thus, many of the required value judgments are not only beyond the understanding of an archivist, they are also beyond their control and purview. Renowned archivist Margaret Child wrote that "the success or failure of the late twentieth-century efforts to preserve our intellectual heritage will be judged by how well what we decide to save meets the needs of the future" (Child, 1992, p. 147). The key word is "decide" and the responsibility for this decision has taken on a different urgency.

Organizations have studied the phenomenon from different approaches across several disciplines. Specific professional groups, such as librarians, with vested interests have focused on their particular areas to inform the overall discussion. Optimistically, there are promising efforts to commercialize long-term preservation as costs for storage decrease. However, policies, procedures and regulations are still insufficiently developed for any unified standard practice. Since civilization crossed the storage production threshold in 2007, the cost for storage may eventually start increasing in response to supply and demand. There will growing need to make, keep, or delete decisions for data. The archive selection decision making process itself will need to change because of the digital transformation. The process needs to evolve quickly as society adapts to the digital modality.

Government organizations are developing digital sustainability guidelines and standards. In 2005, the Library of Congress developed an audit checklist for certifying digital repositories and in November 2011, Consultative Committee for Space Data Systems (CCSDS) issued recommended practices (Council of the Consultative Committee for Space Data Systems, 2011) in an effort to create an ISO standard. In the private sector, researchers at IBM have been concerned with the long-term preservation of data as a part of their business for decades. They store and serve data for major research organizations and corporations worldwide. Even with their methodical, corporate pragmatism, data preservation issues remain problematic. “When a new system is installed, it coexists with the old one for some time, and all files are copied from one system to the other. ...it is hard to predict the cumulative effect that such successive conversions may have on the document” (Raymond, 2001, p. 347).

Walters and McDonald (Walters, 2008) proposed a distributed digital preservation federation similar to the U.S. Federal Reserve Bank regional governance model. When

formalized and implemented, the model would establish institutional trust through its credibility (the perception that professionals are in charge), reliability (outcomes of professionals being in charge), intimacy (acceptance without value judgment) and self-orientation (not self-serving). Building trust models and developing responsibilities across organizations within one country will be a significant challenge. Expanding these models beyond national boundaries creates many additional problems. Governments from technologically advanced countries have also been very interested in preserving their own digital contributions and heritage to society as a whole. In particular, England, France, New Zealand, and Australia recognized the issues of preserving knowledge in its digital form for future generations. These early-adopter countries are developing guidelines to help their nations deal with the issue. Each recognizes that the solution cannot rely on technology alone. The OAIS is one of several international (ISO) standards that attempt to resolve problems in an increasingly connected world.

What Drives Innovation in the Rapidly Developing Digital World

Our ability to pass knowledge and heritage to the next generation is enabled and challenged in ways that have no historical precedent. People will increasingly depend on the digital repositories that contain “officially” processed content and information. Future generations may not have access to the rich, personal letters, such as those between Thomas Jefferson and John Adams, filled with seminal and tangential ideas. The adoption of digital technology is a matter of required behavior in our culture. In the words of one critical observer of technology, Lewis Mumford,

Western society has accepted as unquestionable a technological imperative that is quite as arbitrary as the most primitive taboo: not merely the duty to foster invention and constantly to create technological novelties, but equally the duty to surrender to these novelties unconditionally just because they are offered, without respect to their human consequences (Mumford, 1974, p. 22).

There has been a path of inevitability for digital transformation of society since Turing's 1937 seminal paper in which he wrote about the Turing Machine. "Thus the sequence 001011011101111... and, in fact, any computable sequence is capable of being described in terms of such a table....It is possible to invent a single machine which can be used to compute any computable sequence" (Turing, 1937, p. 241). What he described was the digital, general purpose computer. While Turing's paper describes the world today, it was an abstract theory of a mathematician without any reference point in the world at the time it was written. Early innovators turned his ideas into reality over the next ten years. Many saw a computer as a very powerful calculator. Only a few insightful minds understood the wider use Turing implied. The dawn of the digital world today can probably be traced to the Univac, the successor to Eckert and Mauchly's Eniac computer. (Stern, 1981) The Univac was a general-purpose computer that could be used to solve many different problems. More importantly, it marked the first time that the public saw the power of computers on an election night in 1951. Sperry Rand approached CBS to use its computer to help project election results. That night, as the Univac predicted a landslide victory for President Eisenhower, disbelieving CBS officials refused to air its projected results, which turned out to be less than 1% in error. "Late at night, Collingwood made an embarrassing confession to millions of viewers: Univac had made an accurate prediction hours before, but CBS hadn't aired it...By the 1956 presidential election, all three networks ...were using computer analysis of the results" (Alfred, 2008).

Diffusion of Innovation (DOI) theory (Rogers, 2003) can explain how we can learn to sustain our digital knowledge. The innovation itself must be perceived as an improvement over existing technologies. Applying Rogers' theory, the 1952 election results presented the public with a very observable and stunning demonstration of digital technology's relative advantage for

the future. Early adopters were those who understood the value proposition of the purchase. The relative edge of analyzing huge data sets, heretofore impossible, created a competitive advantage for large corporations and governments. The cost of entry, complexity, and compatibility created significant barriers to everyone else. The technological imperative to improve and disseminate digital technology more widely drove innovation. Sometimes it took years for hardware and software capabilities to meet the demands of forward thinking visionaries. A quarter of a century ahead of its time, a report containing Western Union's 1965 company goals stated:

What is now developing very rapidly is a critical need - as yet not fully perceived - for a new national information utility...(that) will enable subscribers to obtain... the required information flow to facilitate the conduct of business, personal and other affairs" (Union, 1965, p. 3).

Sometimes hardware and software exist, but society does not realize the need. Apple introduced its first "tablet" computer, Newton, in 1993, but didn't succeed until it introduced the iPad fifteen years later. Sometimes, the hardware and software exist only to compete with an organization's core business. Xerox's Palo Alto Research Center (PARC) developed the first graphical user interface (GUI) personal computer, the Alto, in 1973 (PARC, 2012) and essentially gave away its GUI intellectual property to Apple.

The Apple II personal computer was an affordable machine introduced in 1977. As sales increased, people could personally see its benefits, try it, and judge its compatibility with their needs. The interfaces were designed to be more "user-friendly" improving compatibility and decreasing complexity. In 1979, Dan Bricklin introduced Visicalc spreadsheet software for the Apple II that could be used by financial and business communities, large and small. It used the spreadsheet metaphor that is used today and it was free through a new concept called "open-source". The term "killer app" was coined based on the way it energized Apple IIe sales. It is considered the catalyst of the early PC era. (Bricklin, 2012) The pressure to purchase computers

increased because it improved competitiveness for businesses, it was easy to relatively easy to use and inexpensive, improving trialability (Rogers, 2003). A small business that could not afford a mainframe computer in the 1960's could now purchase one of these for a fraction of the cost and accomplish tasks heretofore impossible. Owners could analyze more data, more flexibly creating a competitive advantage to those who adopted. This forced competitors to act quickly or become marginalized. Those that quickly adopted a new digital technology were either richly rewarded or severely punished, dependent on the technology's success or failure.

Around the turn of the twentieth century, data access to residences began to improve significantly. Cable television and telephone companies began distributing TCP/IP traffic into homes. Web access gave the late majority a reason to enter the digital age, perhaps signaling the final transformation to a fundamentally digital world. Mark Weiser (Weiser, 1991) from Xerox PARC is widely considered the father of the ubiquitous computing concept. The term describes the next stage of computing:

My colleagues and I at PARC believe that what we call ubiquitous computing will gradually emerge as the dominant mode of computer access over the next twenty years. Like the personal computer, ubiquitous computing will enable nothing fundamentally new, but by making everything faster and easier to do, with less strain and mental gymnastics; it will transform what is apparently possible (Weiser, 1991, p. 102). Ubiquitous computing suggests a human-machine convergence (Licklider, 1960).

Just as the personal computer made it possible for small businesses and even families to own one, the new model is indicative of the growing dependence on a virtual, digital world. Large companies built and connected much of the global fiber network and solved "the last-mile" problem during the end of the 20th and beginning of the 21st century. The last-mile problem refers to the connection from a neighborhood telephone switch to a household. Telephone companies were incentivized to upgrade antiquated equipment once they saw a business opportunity to provide Internet services. Likewise, cable TV companies upgraded their

equipment to sell network bandwidth to homes and small businesses. Once the network sufficiently penetrated everyday life, a retail company, using the telecommunications industry's terminology, marketed amorphous disk storage called Amazon Elastic Compute-Cloud (EC2) in 2006 (Bezos, 2006). The company purchased disk space to store clients' e-books and discovered that they could sell some of the excess capacity at a competitive price as an unintended consequence. Google introduced its first data center in 2009 in conjunction with the development of its Android operating system and the future it saw in mobile computing. Apple has reinvented and rebranded its online storage offerings three times: “.Mac”, “MobileMe”, and now “iCloud” (Chen, 2011). All three use Mac OS X operating system. OS X is the successor of the Next operating system that Tim Berners Lee first created the World Wide Web. Microsoft launched “Live Folders” in 2007 and rebranded it to “OneDrive” in 2014 (Sanders, 2014). Third party companies dedicated solely to cloud storage such as Box Inc. which launched in 2005 (Rao, 2012) began to further segment the market. By 2015, there were many inexpensive choices and little reason for someone not to have access to all their digital information from any location using one of many types of devices.

Cloud technology assimilates itself into day-to-day life and it can be accessed from a PC, Mac, iPad, or Droid. These multiple entry points into a common virtual framework at an extremely low cost allow the devices and their complexities to disappear into the background. Seventy-five years after Turing's initial paper expressed the idea of a digital world, the technology has begun to fade into the background and our interaction with it have become, to many, as natural as breathing. “The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it” (Weiser, 1991, p. 94).

The Open Archive Information System (OAIS)

NASA recognized the need to manage large volumes of data maintained for projects that had already spanned a generation. They asked the CCSDS to develop a reference model to cope with data from the research of terrestrial and space environmental studies it supported (Lee, 2005). Major space agencies of the world formed the CCSDS in 1982 to offer an opportunity to discuss common problems in the development and operation of space data systems. The model they produced, the OAIS, acts at a highly technical level of abstraction. It became an ISO standard in 2003 (ISO, 2003) and its purpose states that:

“The term ‘Open’ in OAIS is used to imply that ‘Recommendation, as well as future related Recommendations and standards, are developed in open forums, and it does not imply that access to the archive is unrestricted’ ” (Consultative Committee Space Data Systems, 2002, pp. 1-1). Thus, the OAIS is a standard that can be transformed by cultures, laws and norms of those who adopt it. It recognized immediately that digital preservation is not simply about technology. It is about institutional commitment and human behavior. The actual recommendation document states that “An OAIS is an archive consisting of an organization of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community” (Consultative Committee Space Data Systems, pp. 1-1).

The reference model has been adopted far outside of its initial purview.

OAIS-compliance has been a stated fundamental design requirement for major digital preservation and repository development efforts at the U.S. National Archives (NARA), U.S. Library of Congress (LC), British Library, National Library of France (BnF), National Library of the Netherlands (KB) the Digital Curation Centre (DCC) in the UK, Online Computer Library Center (OCLC) the JSTOR (Journal Storage) scholarly journal archive, as well as several university library systems and space agencies (Lee, 2005, p. 4).

The OAIS constitutes the framework for the new digital environment. The information package contained in the OAIS carries with it attitudes and beliefs of the authors who helped create it, as well as the archivists and administrators who decided to maintain it. Intentionally and unintentionally, generational and cultural biases are introduced to a file each time the archivists attend to it, potentially preserving the object forever in the present tense. Digital sustainability practices require that the package be revisited every five to ten years by archivists. Information package retention becomes a part of the administrative resource allocation process. Each time a package is evaluated based on the standards of the present with unknown cumulative effects. When discarded, the reduced number of copies inherent in the digitized world makes it less likely that the information will be retrieved in the future from a rediscovered source. Unlike written text, which can sustain generations of ambivalence and inattention, digital data requires regular attention either by a person or an organization, to maintain it. A single break in the chain, in Conway's words, permanently loses a part of culture and history. The OAIS provides

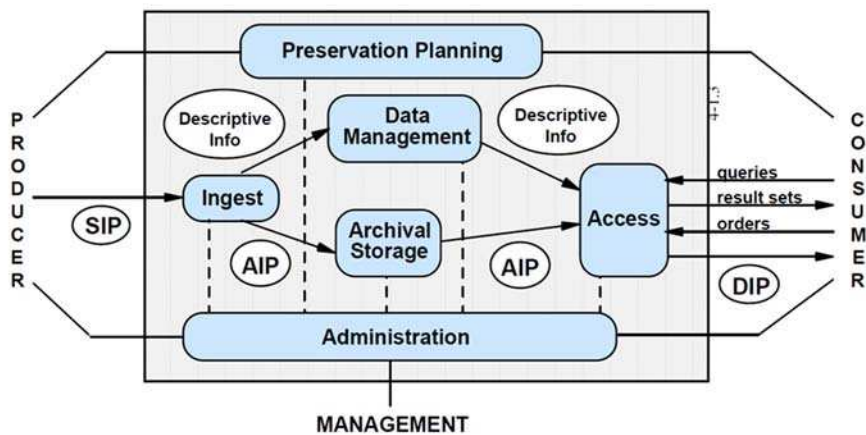


Figure 2 OAIS Model(Council of the Consultative Committee for Space Data Systems, 2011, pp. 4-1).

an understanding of where and why resources need to be allocated to preserve these documents.

Library of Congress' Seven Digital Sustainability Factors

The Library of Congress uses seven digital sustainability factors that help with its decision-making process. “These factors influence the likely feasibility and cost of preserving the information content in the face of future change in the technological environment in which users and archiving institutions operate” (Library of Congress, 2007a).

Table 1 Library of Congress' Seven Digital Sustainability Factors

Disclosure	Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating digital content.
Adoption	Degree to which the format is already used by the primary creators, disseminators, or users of information resources.
Transparency	Degree to which the digital representation is open to direct analysis with basic tools, including human readability using a text-only editor.
Self-Documentation	Degree to which the metadata is descriptive of the digital object.
External Dependencies	Degree to which a particular format depends on particular hardware, operating system or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.
Impact of patents	Patents inhibit the ability of archival institutions to sustain content in that format.
Technical Protection Mechanisms	Degree to which content may be replicated on new media, migrated and normalized in the face of changing technology and disseminate it to users at a resolution consistent with network bandwidth constraints.

Note: (Library of Congress, 2007b)

These factors help archivists to bridge the differences between the written and digital world. They focus our attention on the essential foundations of archives in order to preserve knowledge. In reality, the factors can be applied to written records and are either exacerbated or improved as they are digitized. For example, copyright is a critical element of any business model for publications. It has protected and rewarded creative people for over one hundred years. It is codified in our legal system and incentivizes invention. The transition to a digital world is forcing the community to rethink ownership rights and control. Copyright hinders adoption of community-based archives such as the Digital Commons. It also hinders submissions to local digital repositories. The information in Digital Commons or digital repositories is typically open, free, and discoverable. This places these systems in direct conflict with copyright and its profit. Even if a manuscript is copyright protected such that total free access would not be available when a file is submitted, it would be difficult to assure the document will not find its way onto a free, widely accessible source. There is a tension between archivists who preserve knowledge, businesses that want to control information for profit, and a government that needs to manage its empire. Experts will need to thoroughly vet the transition from written to digital records using the seven sustainability factors to wisely archive knowledge.

The National Institute of Health Data Sharing Policy

Two of the significant advantages digital sources have relative to paper are portability and replicability. They are more “space bound” (Innis, 1972). The digital model is challenging rules and mores such as copyright and academic honesty as previously discussed. There is a tremendous clash between those who advocate open access and ownership. US government-funded research is paid for by the public, therefore deemed to be the property of the public. Beginning in October 2003, the National Institute of Health implemented a policy describing

scientists' responsibility to share final research data acquired during activities sponsored by the NIH. "Starting with the October 1, 2003 receipt date, investigators submitting an NIH application seeking \$500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible" (National Institutes of Health, 2003). This requirement implied that researchers would have to give some consideration of their data preservation before starting a research project in considering how and where they would store it to make it accessible. In order to make it widely accessible, they would need to use relatively standard software packages. The factors that make it more accessible, in this case, creating a more sustainable file. It also implies organizational responsibility that would be required to invest in infrastructure (capital investment and support personnel) in order to support the research being conducted on their behalf. At some point in time, this requirement should stimulate dialogue throughout the organization.

The policy was a limited, but important, first step in promoting digitally sustainable practices (it was directed at NIH grants of \$500,000 or more). It defined the timeframe that it had to be accomplished: "NIH expects the timely release and sharing of data to be no later than the acceptance for publication of the main findings from the final dataset" (National Institutes of Health, 2003). The NIH held workshops in support of the requirement, introduced ideas to address proprietary data, and described methods to share data.

The National Science Foundation Data Plan

On January 18, 2011, the National Science Foundation (NSF) instituted a requirement that all researchers complete a simple two-page data management plan (DMP) supplement for each grant submission (National Science Foundation, 2012b para 2). NSF grant submissions are not accepted without one after that date. If the research project expects no data, then the DMP

may state that no computer data are to be generated in the research. The mandate's stated goal is to make research widely available to enhance discovery. The NSF DMP requirement is quite brief (See Appendix E to read the entire requirement) and provides only general guidance for the DMP. "This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results ...and may include" (*the following guidelines in table 2*):

Table 2 National Science Foundation's Guidelines for Inclusion in a Data Management Plan

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

Note: (National Science Foundation, 2012b)

The Foundation states that it will rely on the various Directorates, Offices, Divisions, Programs, or other NSF units to refine the requirement in accordance with each domain. "What constitutes such data will be determined by the community of interest through the process of peer review and program management" (National Science Foundation, 2012a). As of March 27, 2012, six of the seven research directorates (Biological Sciences, Computer & Information

Sciences & Engineering, Education & Human Resources, Engineering, Geosciences and Social, Behavioral and Economic Sciences (SBE)) have given Directorate-wide guidance and implementation plans. The five divisions from the Mathematical and Physical Sciences Directorate have separate guidance policies with slightly different wording. An example of the rewording can be seen by comparing the definition of “data”. The NSF definition:

What constitutes such data will be determined by the community of interest through the process of peer review and program management. This may include, but is not limited to: data, publications, samples, physical collections, software and models”(National Science Foundation, 2012a).

The Directorate of Biological Sciences (BIO), citing OMB Circular A-110 (Office of Management and Budget, 1999) defines data as “...the recorded factual material commonly accepted in the scientific community as necessary to validate research findings.” This definition includes “both original data (observations, measurements etc.) as well as metadata (e.g., experimental protocols, software code for statistical analysis etc.)” (National Science Foundation, 2011a). The Social, Behavioral and Economic Sciences Directorate (SBE) elaborates even further:

Research data are defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This "recorded" material excludes physical objects (e.g., laboratory samples). Research data also do not include:

(A) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and

(B) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study (National Science Foundation, 2011c, p. 3).

Each of the twelve Directorate/Division data management plan requirements recaps NSF general guidelines, but includes further elaboration and implementation guidelines. For example,

where the NSF does not discuss how the policy will be put into practice, the Social, Behavioral and Economic Sciences Directorate explains in more detail:

The DMP will be considered by NSF and its reviewers during the proposal review process. Strategies and eventual compliance with the proposed DMP will be evaluated not only by proposal peer review, but also through project monitoring by NSF program officers, by Committees of Visitors, and by the National Science Board” (National Science Foundation, 2011c, p. 4).

NSF general guidance does not discuss follow up, but the Education and Human Resources (EHR) Directorate explains: “After an award is made, data management will be monitored primarily through the normal Annual and Final Report process and through evaluation of subsequent proposals” (National Science Foundation, 2011b, p. 2). Some standardization among the Directorates may occur in the future as the NSF reviews its policies to develop best trans-disciplinary practices.

The intent of the NSF data-management plan is to require researchers to share their information in order to spur invention and reduce its cost for research. In so doing, the NSF enhances digital sustainability activities by compelling researchers, administrators, and technicians to develop and accept standard practices, file formats, and workflows. The NSF data-management plan can also stimulate organizational resource commitment and policy changes in order to support digital research.

Council on Library and Information Resources pub154

The Council on Library and Information Resources (CLIR) released a web report “The Problem of Data” (Spencer, 2012) in August 2012 focusing on the data management practices of university researchers. Spencer’s report focused on five institutions in northeastern United States using ethnographic interviews with faculty, postdoctoral fellows, and graduate students in several social sciences disciplines.

The goals of the study were to identify barriers to data curation, to recognize unmet researcher needs within the university environment, and to gain a holistic understanding of the workflows involved in the creation, management, and preservation of research data (Spencer, 2012, p. 3).

The study reveals the short-sighted attitudes of most researchers. Among the findings of the study are that researchers are dissatisfied with their own level of expertise, but that few are thinking about long-term preservation. They cite that the demand of publishing undermines efforts to change their behavior unless it helps them to complete their research. The organization itself has the responsibility for providing the policies, the resources, and fostering attitudes to instill sustainable digital behaviors. Administrators need to have the providence to create systems that support their research faculty.

Digital technology has been available for over half a century. The widespread use of digital technology in business is about twenty-five years old – since the advent of the World Wide Web. It is only in the past decade that institutions have begun to formalize how our knowledge for future generations will be preserved. Each of the aforementioned efforts is part of an iterative process. Researchers develop digital preservation practices for their research with constraints imposed by the medium, assistance from exemplars and peers, and with direction from granting agencies. They can apply these practices to their personal lives and provide feedback to the process based on these experiences, making small changes to the system to fit their needs. Thus, the digital environment is a creation of human activity that is both technical and social. Solutions require improved understanding of the communication that crosses organizations and that recognize individual needs. The efforts at this early stage of the paradigm shift will influence future generations' perceptions of who we were, what was done, and why.

Changing Researcher Workflows for the Digital Age

There is a need to understand researcher workflows with respect to digital self-archiving attitudes and behaviors in order to improve digital sustainability. Access to local expertise and resources can improve their experiential knowledge. Support and guidance from campus leadership will improve the dialogue on campus that should provide increased collaborative solutions. Researcher behavior is strongly influenced by their domain. Peers in the field are very important to research practices. Acts, such as applying for grants to NSF or NIH, influence overall researcher knowledge, perception, and attitude. However, a researcher's job appointment is within a department in a college or university to which they are accountable. Researchers, technologists and administration should have a dialogue on campus to resolve and learn best practices. This is achieved on a campus through motivation and leadership.

In the US, both the public and private sector have begun to develop the framework, policies and practices to address the issue. CCSDS created OAIS to preserve archived data, but it also provides a roadmap for the new digital ecosystem. It offers a framework that can be used to identify roles and responsibilities within a bounded system. The Library of Congress is developing specific guidance for preserving digital artifacts. Both the NIH and the NSF have created mandates that require the consideration of an all-inclusive data management system by researchers, administrators, technologists, and organizations. NIH and NSF requirements are important events in the trend to encapsulate and maintain humanity's knowledge digitally. Researchers who apply for these grants must consider the future access of their data, in so doing should consult expertise within their organization and their domain. Local events such as data preservation workshops and classes provide another communication channel to increase dialog and reinforce messaging. Local resources and support are also critical to adoption and

adaptation to a sustainable digital life. Hardware, software and preservation expertise enable researchers to comply with new policies. The digital world is deceptively insubstantial and relies on complex interrelated activity whose reliability hasn't been tested by time.

Theoretical models

Socializing The Digital World

The question “How can organizational resources be effectively communicated to researchers to improve their file management skills?” is complex and rooted in the digital transformation itself. The transformation is driven by the motivations of McLuhan’s “space” over “time” (McLuhan, 1964). As described in the literature, it is a fundamentally different paradigm in many ways. Written text can withstand long periods of abandonment with little intervention. We expect that it can be retrieved when we need it. Behaviors have been built around this “time bound” characteristic over centuries. We have experienced few notable disasters such as the destruction of the Library of Alexandria in our history. We have only recently reached a point in time that access of digital data twenty years old has become an issue. We are just beginning to realize some of the consequences of the transformation.

Over forty years ago McLuhan stated “Today after more than a century of electric technology, (we) have extended our central nervous system itself in a global embrace, abolishing both space and time as far as our planet is concerned” (McLuhan, 1964, p. 3). At the time, the personal computer was almost twenty years into the future and the World Wide Web would not be invented for three decades. A half a century later, we may finally be realizing this vision. If we are to live in this world, there is an imperative to safely store the digital data to sustain it for the future.

As we place our knowledge into digital repositories, we become more dependent on their reliability and accuracy. The change from written to digital is as significant as the transition from oral to the written tradition. The digital paradigm improves speed and access to information at the expense of preservation. The transition to digital methods and practices in the

past three decades created deep, fundamental changes in our way of life. The evolution has been extremely dynamic over the past thirty years as developers and users reinvented how to build and use the new technologies more than once. While new tools are fundamental to societal development, they typically are “socialized” and standardized in accordance with the group’s acceptance. Philosopher Jacques Ellul (1964) observed that the human purpose of technology is replaced by a utilitarian view of the task at hand. In the 1880’s “Sewing machines were decorated with cast iron flowers...the machine can become precise only to the degree that its design is elaborated....in accordance with use... Abstract techniques and their relation to morals underwent the same evolution” (Ellul, 1964, p. 73). Computers are disappearing into the cloud altogether. What remains is a device to access the virtual space. Since the introduction of the PC thirty-five years ago, data storage evolved rapidly from 5 ¼ inch to 3 ½ inch floppy, from CD’s to DVD’s to Blu-Ray DVD, and to external hard drives. We are reaching the stage of network storage, where it may be possible to assume that media formats will stabilize. However, standard media formats are only a part of the problem. Researchers can now choose from many cloud storage options: Box.com, Dropbox, Google Drive, Microsoft OneDrive, and Amazon Prime. Each of these has its own risk. Each option uses a license agreement that declares the rights and responsibilities for the company and the client. The license agreements can be changed quickly to the advantage of the vendor, potentially jeopardizing some data stored on it. The OAIS model can provide a framework for organizational and individual policies and procedures to ensure the longevity of digital artifacts in this dynamic environment.

Our ability to pass knowledge and heritage to the next generation is being challenged in ways that have no historical precedent. Simply organizing and preserving the data becomes an issue. We will depend more on the digital repositories that contain “officially” sanctioned

content and information. Future generations may not have access to the rich, personal letters filled with information not initially considered important enough to save. Moore's Law is a term used to define the speed and growth of technology. In his seminal paper in 1965, with four data points, he observed that "The complexity for minimum component costs has increased at a rate of roughly a factor of two per year" (Moore, 1965, p. 115) Ten years later, he revised his projection, stating that "the new slope might approximate a doubling every two years, rather than every year." (Moore, 1975, p. 12). However, personal and organizational behaviors and responsibilities have lagged behind. Digital preservation is an act by people, either individually or in groups. Digital data will become more "socialized" as it is integrated with the networks of it is connected to.

Diffusion In The Rapidly Changing Digital World

Diffusion of Innovation (DOI) theory (Rogers, 2003) can be used to help shape our understanding of how we can learn to sustain our digital knowledge. The theory explains how OAIS has become the accepted model to describe digital artifacts long-term survival in a data ecosystem. Lee (2005) discussed the adoption and diffusion of the OAIS standard using Roger's DOI model. Lee states that OAIS implementation is a good candidate for future research. Implementing the model includes adopting or adapting several components that include policies, standards, resources and the technology itself. In his seminal work, Everett Rogers (2003) defined diffusion as a process in which innovation is communicated within a social system. It is a mature theory that is rich with methodology including case studies, policy analysis, network analysis, surveys, and experiments. However, Rogers also recommended that "Diffusion scholars should seek alternatives to using individuals as their sole units of analysis" (Rogers, 2003, p. 125) to overcome what he named the individual-blame bias. The problem he describes

is that the model indicates the success or failure of the individual within the system rather than the success or failure of the system.

Rogers first published Diffusion of Innovations in 1962 articulating his theory on the role that communication plays with the speed that societies adopt new technologies and ideas. Several complex factors, or barriers, govern adoption rates for any innovation. The innovation itself must be perceived as an improvement over existing technologies. He called this its relative advantage. If there is great improvement, it is more likely that there would be rapid societal adoption. Even if there is a relative advantage, any one of several factors may inhibit or even stop adoption. An innovation that does not conform to cultural values and traditions of a local culture whether it is tribal, national or corporate is unlikely to be accepted. He cites examples that include water boiling in a Peruvian village and Xerox corporate accepting Xerox PARC's personal computer in the early 1970's. In the first case, villagers linked boiled water to illness; therefore, inhabitants learn that boiled water is bad. In the second example, the personal computer failed at Xerox since it competed with its core business. Individuals need to determine whether there is a relative advantage by testing the new technology in their own environment. If trialability, as Rogers calls it, is limited, the pace of adoption will slow. Marketing campaigns include free trials to reduce this barrier to the public and increase diffusion. Diffusion of Innovation Theory advanced archetype personalities and adopter categories in his theory to explain individual and organizational adoption to any innovation. Early adopters are said to be opinion leaders who typically introduce innovations to their group. These opinion leaders have what Rogers called "heterophily", or a perspective that allows them to bring ideas in from outside the group. These relationships are critical to introducing new information to the group which Rogers called a "difference in matter-energy"(Rogers, 2003, p. 3). Grannoveter (1973)

studied the exploitation of this relationship by trying to understand weak ties between individuals who exist in different groups. Using Milgram's small-world experiments (1967) and Rogers Diffusion theory, he proposed that by finding individuals who are weakly tied to groups, one could find opinion leaders and more effectively leverage communications. Watts and Strogatz (Watts, 2003) later confirmed his ideas empirically through a quantitative approach using network modelling. This approach showed that a small-world phenomenon occurs in highly clustered networks with short path lengths. As people form cliques, communications channels are focused within the group. Opinion leaders connect to other cliques for new information. The result of these factors is higher clustering and fewer steps to other groups thereby creating a more highly connected network.

The transition to a digital world is a transformative innovation comparable to the Industrial Revolution. Just as the factory was integral with the industrial revolution, it is only one component. Industrialization needed an educated urban population, tools, laws and process to be successful. Not all concepts from DOI apply directly to the transformation as they would to a discrete innovation. It is easy to conflate the invention of the computer with the creation of the digital world. Understanding the digital environment that has been built requires a broad perspective. If it can be done, it will be up to historians far in the future to look at the timeline of the twentieth and twenty-first centuries to intelligently discuss the knowledge, persuasion, decision, implementation and confirmation stages of the digital transformation.

There are some DOI concepts which can be discussed, such as relative advantage, compatibility, complexity, trialability, and observability. Digital assets have a tremendous relative advantage of speed and portability to convey information. Digitization gives greater access to information, which makes it compatible to individuals under most circumstances. It is

hard to understate the relative advantage of digital information. The information digital data encodes and decodes includes facts about people, places, and things. The advent of 3D printers means we can create tools on the space station, reducing the room needed for the large number of specialized tools or DNA in a lab to reconstruct our failing body parts from digital information. We can now print an automobile (Harrop, 2015), which begins to call into question the very existence of industrial factories.

The relative advantage of digital things far outweighs the complexity for the few early adopters. Speed of information allows private organizations and governments to shorten decision cycles and to control their enterprise. One example of the increased pace is average daily volume on the New York Stock Exchange. Each trade is the outcome of a decision based on information from multiple sources in a highly competitive environment. As the speed this information can reach traders increases, decisions can be made more quickly. In 1965, six million shares were traded. In 1985, 109 million shares were traded, and by 2005, 1.6 billion shares were traded (New York Stock Exchange, 2015). The trading environment is a microcosm of the space-bound world McLuhan and Innis described. Traders live on the edge, willing to dedicate vast resources to experiment with innovations as a matter of survival. There are financial incentives to simplify the systems, processes, and tools for acceptance within the organizations. The digital paradigm's speed and accessibility mean these improvements quickly spread to all but the most recalcitrant late adopters to negotiate the digital frontier.

Open source is a new concept that gives new meaning to trialability and observability. It is part of the growing sharing economy. This part of the digital world allows anyone to have powerful software and services free. It not only allows, but expects people to improve their creations. Feedback between customer and creator can be near real-time. Participation in the

development is welcome. Digital asset access, portability, short innovation cycles, and customizability continue to drive society to adopt and adapt to the digital age rapidly.

Digitally sustainable practices for data are both individual and organizational responsibilities in which an information package is the result of a complex system of behaviors, decisions, policies, and practices. According to Rogers, the success of any strategy within the system depends on how behaviors are shared. Public opinion is less bound to location in a virtual world. Finding group similarities beyond political boundaries and across cultures is possible and incentivized by the market. In accordance with diffusion theory, decisions may be optional, collective, or authority driven (Rogers, 2003, p. 28). The National Science Foundation is imposing an authority innovation decision on researchers who may act within their discipline to create a collective decision driven by collegial peer pressure while working at a university where academic freedom mandates that the requirements be optional. There are several roles and responsibilities with respect to digital preservation in an organization. Senior administrators need to understand the strategic, long-term importance of the institution's digital assets. Their choices create the options available throughout the organization. Technologists need to maintain the systems. The archivists are the caretakers and the exemplars that practice and create practices for others to emulate. They must not only know the technology, but law and policy for all the entities that the archives are intended for (journals, government studies, privately funded organizations). Their behavior should be strongly influenced by their field, discipline, associates, and accepted social behavior. The authors are ultimately responsible for ensuring archives are managed both locally and on a remote system (at least initially). There should at least be an explanation of how these aspects fit together to ensure that information stored today will be available twenty five years from now, as well as two hundred fifty.

According to Rogers, the success of any strategy within the system depends on how decisions are shared. Digitally sustainable practices are both individual and organizational in that an information package is the result of complex interactions between behaviors, policy and resource. Innovators may make optional, collective, or authority driven decisions (Rogers, 2003). For example, the NIH in 2003, then NSF in 2011, and as of 2013, most Federal government agencies have imposed authority-driven innovation decisions on researchers in the form of a data management plan requirement to increase access to the funded research in which they are investing. In response, researchers contribute rules and practices of their own discipline to feedback via implementation decisions from their university campuses.

Technologists build and maintain the systems that store information. The range of systems includes a safe desktop environment (protected from malware on standard software) on which researchers work to integrate cloud-based solutions that have rapidly grown since their inception in 2006. IT specialists balance security and protection with open access in an open-ended collaborative environment. Archivists are the caretakers whose duty includes creating digitally sustainable practices. They need to understand each document within the context of the field or discipline, as well as the policies, practices, and laws that a document may be subject to (i.e. journal, government, and privately funded organizations). IT specialists and archivists, together, contribute to evaluating the effect hardware and software upgrades will have on documents to minimize data loss. Researchers themselves operate between the organization of the university system (department head, dean, or provost) and the organization of their discipline (peers and colleagues) globally. Their affiliation networks inform their digital management decisions that are typically dependent on local file systems. Authors are ultimately responsible for decisions about their archive as it passes through stages of a local file to a published

document. They are responsible for taking the initial, critical steps so that their digital artifacts stored today will be available twenty-five years from now, as well as two hundred fifty years from now.

The Act Of Making And Socializing Our Tools

Activity Theory is a descriptive meta-theory with the premise that tool production results when individuals engage with their environment. “Tools are, in fact, “exteriorized” forms of our mental processes shaped by culture, history, rules and other variables” (Morten Fjeld 2002, p. 153). It shares similarities with Suchman’s work. Suchman states “Cognitive scientists today maintain the basic premise of de la Mettrie with respect to mind, contending that, mind is best viewed as neither substantial nor insubstantial, but as an abstractable structure implementable in any number of possible physical substrates” (Suchman, 1987, p. 7). Our interaction and mutual evolution with our tools is part of our shared cognitive experience. While she uses a cognitive approach, AT attempts to account for community, the economy, the environment and the rules that dictate what we can and cannot do. Alexander Luria, Alexei Leont’ev and Lev Vygotsky developed AT in the early twentieth century in the former Soviet Union (Engeström et al., 1999). The approach emerged from the Moscow Institute of Psychology and gained more recognition in the West with the end of the Cold War, the emergence of Human-Computer Interface (HCI) and the energy of expatriates in the US such as Bedny (Wilson, 2008). Figure 3 depicts Engeström’s representation of AT.

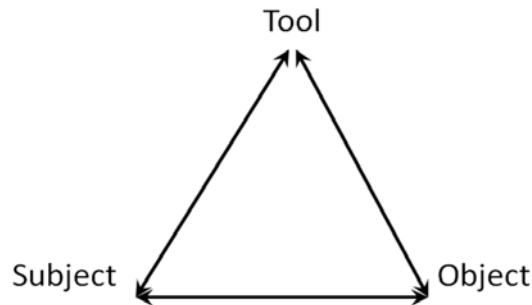


Figure 3 Engeström's representation of Vygotsky's Activity (Engeström, Miettinen, & Punamäki-Gitai, 1999, p. 30).

Activity theory stresses the development of cognition as a unity of biological development, cultural development, and individual development. It has a strong ecological and functional-historical orientation. It also stresses the activity of the subject and the object orientation of this activity (Hjørland, 1997, p. 80).

AT evolved since first introduced in the former Soviet Union and has champions in the Scandinavian countries and is particularly strong in Finland. A principal of AT is “the unity of consciousness and activity” in which the mind emerges in evolution through activity in relation to the external environment. A fundamental form of human activity is external activity with practical goals. The focus on the activity and de-emphasis of the subject and object provides an alternative perspective for systems development as well as social norms. “The mind is a special ‘organ’ that appears in the process of evolution to help organisms to survive. Thus, it can be analyzed and understood only within the context of activity” (Nardi, 1996, p. 107). The core idea expands the way we conceptualize our interaction with the world. By adding the instrument or tool, Vygotsky both recognizes its importance and implies a social cognitive element. Agency is located in the activity while the subject, object and tool are all transformed in some way by the process. As these artifacts become available to the public, the tool, the individual using it and society change as behaviors focus on the adoption of the activity. As we create tools (both physical and virtual) to interact and engage our environment, those tools become social artifacts.

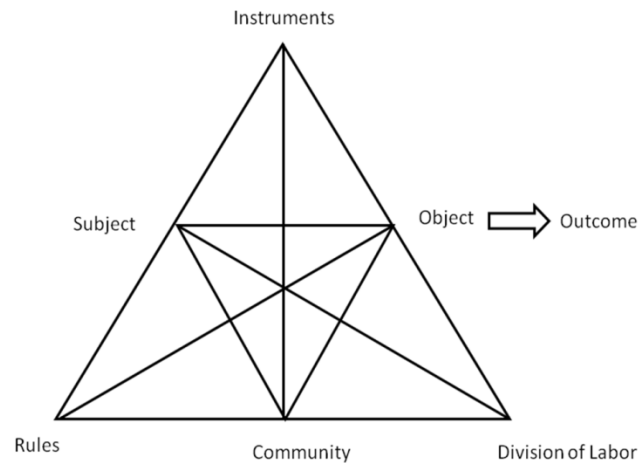


Figure 4 Engeström enhancement of Vygotsky's AT representation (Engeström et al., p. 31).

Wilson (2008) sought to explain information-seeking behavior as part of a process of internalization/externalization that individuals go through to adapt their mental processes with the tools we create. Bedny (Chebykin, Bedny, & Karwowski, 2008) has written extensively on AT as a structured system from a Human-Computer Interface perspective. Bedny, et al. explored a variant of AT called Systemic-structural Activity Theory (SSAT). SSAT "...demonstrates that learning can be viewed as an active regulative process, and strategies of performance can be described based on analysis of self-regulation mechanisms"(Chebykin et al., 2008, p. 46). In contrast, a group from Scandinavia have taken a different approach to AT. The focus is shifted to the "historically located" activity as the fundamental unit of analysis. "While featuring the crucial link between subject and object, this approach features the essentially social nature of activity and the centrality to it of durable cultural artifacts" (Sannino, Daniels, & Gutierrez, 2009, p. 29). Cultural-Historical Activity Theory (CHAT) is a means to apply the theory in context of society. "The visual representation of the triangle was a way to condense and convey theory in research collaborations with practitioners...(It was) designed to destroy the myth of directness in learning and teaching" (Sannino et al., 2009, p. 13). The top third of the model constitutes Vygotsky's Foundation of subject-object-tool. All interact within an 'activity'

transforming each other based on their innate advantages and limitations. Engeström expanded the activity to include rules, community, division of labor and all interactions in order to represent the social elements of an activity system. His third generational model (Figure 4) looks at the joint activity as the unit of analysis rather than individual activity. In this model he is interested in the process and its impact on social transformation.

Model and Research questions

The essential elements of the OAIS can use the AT model to analyze various parts of the digital preservation environment holistically. Figure 5 provides a conceptual framework to analyze the creation of a SIP, the input to a professional managed archive, from a research

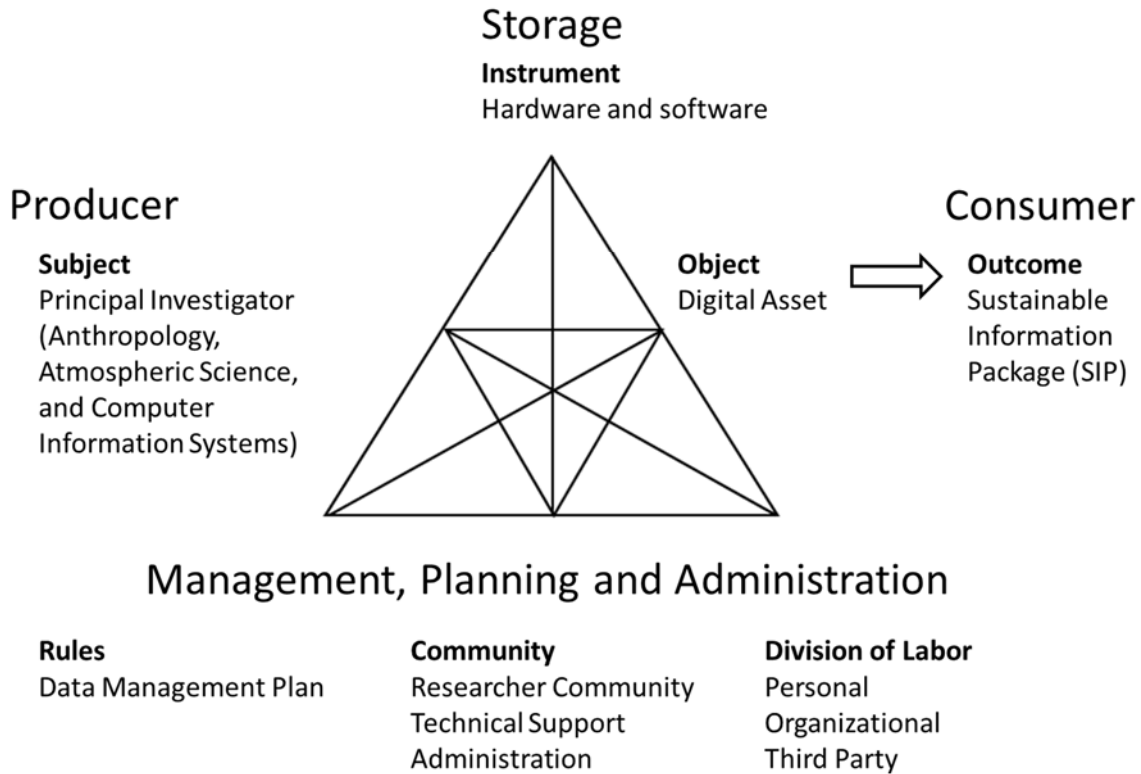


Figure 5 Proposed transformation of research data to a sustainable information package within Activity Theory construct.

activity using Engeström’s model. The subject (Producer) works with available instruments (Storage) along with rules, community and labor (Management, Planning and Administration) to produce an outcome from an object (Object) for the consumer, who in this case would be an archive. Within the research model, multiple methods are used to analyze stimuli and behavior to improve diffusion of digital preservation practices at the university.

Using this model, the researcher sought specific answers to information, events and policies about the way research data is maintained. Each of these questions then relate to the model.

R1 – How do researchers manage digital data at a major research university?

R1a – How much data do researchers have to manage?

R1b – What storage resources do researchers use?

R1c – What digital data management training activities do researchers attend?

R1d – What digital file management practices do researchers use?

R1fe– What data do researchers expect to share from their research?

R1f – What critical research data loss events have occurred?

R1g – To what extent have researchers needed to file data management plans?

R2 – How can digital preservation be communicated to researchers to improve the permanency of their data?

CHAPTER 3 METHODS

Synopsis

The case examines how rules, resources, and training can be communicated to research faculty to raise overall awareness and improve digital preservation in the conduct of research. A goal of this study was to find an approach to influence better digital preservation at the institution through communication with the knowledge that faculty are strongly influenced within their discipline. Since faculties are evaluated, supported, and tenured at the department level, messages tailored at the department level should be more successful than if from the college or university level. In order to understand how units interact with each other with respect to data management, the study asked how researchers in the three departments manage their own digital data. Individual researchers are the unit of analysis.

The approach is to use the OAIS and the Activity theory to represent a construct of the interrelated processes that describe the digital preservation process. The researcher used multiple data collection methods that included institutional data, multiple transcribed interviews, and face-to-face surveys to create a rich data set. The study was approved by the IRB on July 2, 2013, protocol number 13-4247H. The study group is faculty from three selected academic departments at Colorado State University (CSU). The academic units were chosen for their diverse environments. The focus is individual digital data preservation practices in the departments while conducting research. Each research faculty is the first critical steward of digital data. They establish the value of the data and determine its initial disposition. The researcher focused on successful NSF grant submissions since its data management plan condition requires researchers to explain their data management process and it represents almost

10% of total institutional grant funding. Researchers are developing new skills and workflows while standards are still emerging, which makes submissions more difficult. Expertise and resources exist on campus which can facilitate researcher compliance. The university library provides an institutional repository and training. Information technologists maintain storage for file systems and data backup. Communicating resource availability to research faculty is important for informed and swift adoption, which is an imperative in the rapidly evolving and growing digital environment. It is important to understand the network of interested parties who can exchange digital management best practices. Social Network Analysis (SNA) is a powerful tool that can reveal important relationships and characteristics of the institution's research system.

SNA provides a systems-level approach (Scott, 2000) to analyze how people and groups relate to each other. It can account for the diverse interactions as people, tools, rules, resources, and communities develop a virtual ecosystem, hopefully, into a sustainable one. Scott (2000) states that social network analysis describes and measures relations between people, objects, or concepts instead of the attributes of entities. It is a relational model with its own statistics, not the attributional (mean, mode, distribution) that social scientists frequently use. It disregards agency and subjectivity. Its statistics include such measurements as density, betweenness, closeness, centrality, structural equivalence, and structural cohesion. An individual's network

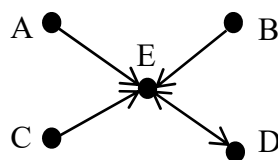


Figure 6 Sociometric Star.

position and their links to those around them are the keys to their importance instead of their

individual attitudes and beliefs. Network analysis developed, according to Scott, “from the structural concerns of the great anthropologist Radcliffe-Brown” (2000, p. 4) who, in the 1920’s, analyzed social organizations in the Andaman Islands and later among the aborigines of Australia. His work influenced sociometry research developments in the 1930’s by Jacob Moreno. Moreno’s research led to the development of the sociogram, which Scott describes as a precursor to network analysis research. His sociometric star represents the individual as a point or node, relations as lines and arrows showing directionality. Notice in figure 6 “A”, “B”, “C” and “D” like “E” but “E” only likes “D”. Lewin combined Gestalt concepts with this simple representation to develop Field Theory. He posited that a social group exists in a social space, or field. He used mathematical techniques such as topology and set theory to understand the structure of the group. Additionally, as depicted in figure 7, each line can have direction (“A” works for “C”) depicted by the arrow, can have intensity (“B” works for and sits on two committees with “C”) depicted by the multiple lines and can be signed (+ or -) to indicate positive (“A” likes “C”) or negative (“D” does not like “E”). The aim of field theory is to explore, in mathematical terms, the interdependence between group and environment in a system

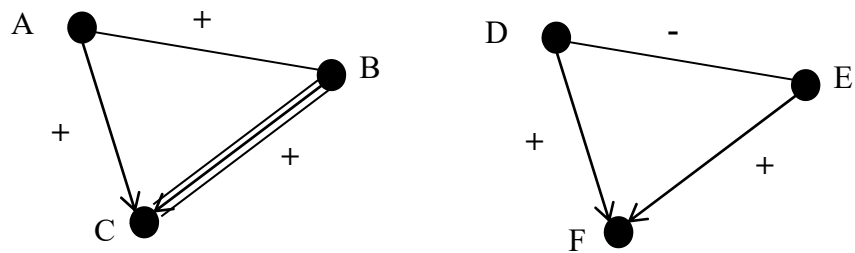


Figure 7 Balanced and unbalanced structures with multiple connections between “B” and “C”.

of relations” (Scott, 2000, p. 11). Heider used positive and negative signs to represent either positive or negative affinity. Social systems and those with mixed signage were out of balance,

and naturally trend towards a balanced system. “D” does not like “E” in figure 20 putting the system “out of balance”.

Scott cites Harary and Norman’s mathematical work as a major breakthrough in the 1950’s, “This breakthrough consisted of moving from the concept of cognitive balance in individual minds to that of interpersonal balance in social groups” (Scott, 2000, p. 12). Milgram’s research into our connectedness led to his small world experiments in the 1960’s. In the experiments, he demonstrated that one could reach anyone in the world within six steps when he recognized that if the average person knows one hundred friends, the network would be 10,000,000,000 people in five steps. Although it does not take into account multiple common connections in the network, it demonstrates how quickly a network can grow. Granovetter (1973) used the results of Milgram’s experiment to propose the “strength of weak ties” theory. Weak ties represent the connections between different cliques that transmit information at great distances across a network. Cliques are represented by strong ties within an association. Weak ties are represented by individuals who are members of a professional association, but also have contact with another group to which their association has limited access. Granovetter proposed that it is through these associations that information is most effectively transmitted to distant networks. The term “distant” in this case should not be interpreted as physical distance but network distance. In fact, two networks may be very close physically, but distant socially. This concept is critical in an academic setting. An English professor may be located down the hall from a physics professor, but actually work more closely with a colleague on a different continent.

The increasing power of and accessibility of computers allowed researchers to pursue more complex models with larger datasets. The increasing power of network analysis tools

allowed researchers to apply its statistics to social sciences, as well as hard sciences such as physics. Network nodes, after all, can be people, objects, or concepts. Watts published research (2003) in which he confirmed Granovetter's assertion that the social network is held together by the strength of weak ties. After the turn of the 21st century, network analysis has exploded as Google worked to improve the quality of its marketing data, security agencies attempted to develop intelligence about terrorist cells, health agencies tried to stop epidemics, and scientists wanted to better understand our natural world.

Connected Research - Institutional Records

The researcher downloaded all available institutional records that included 44,681 successfully submitted grant proposal records spanning twenty-eight years from university data available from the Vice President for Research website (Colorado State University Vice President for Research, 2015) for the years 1987 – 2014. A successfully submitted grant is one that received an award. Given that only awarded grants are included, in-degree and out-degree are equal. Each record contained the principal investigator's name, administering department code, college, sponsor, title, award date, fiscal year, and award amount. The research linked each grant research project by individual researchers with their respective sponsor using Gephi data visualization software (Bastian, 2009). Nodes and connections were solely weighted by the number of connections in a dyad. The researcher downloaded and combined comma-delimited files into an Access database, associated each researcher to their respective college and imported the data into a Gephi project workspace.

Initial Interviews – Create Survey Categories

The researcher conducted interviews in the offices of two faculty members from each of three departments for a total of six interviews. Each interview session lasted approximately one

hour. The interviews were recorded and transcribed. The questions were based on the CLIR Pub 154 and its survey (Appendix A) (Spencer, 2012). The researcher chose one relatively new faculty member from each department and one more experienced. The researcher framed the open-ended questions around a recent project they had worked on. The researcher grouped the answers into broad categories: data management training, data management practices during and after the project, collaboration during the project, preservation activities at the conclusion, and specific events that may influence the way they work with their data. The researcher used the NCT (Notice things, Collect things, Thinking about things) model (Friese, 2011, p. 12) with Atlas.ti qualitative analysis software to derive themes from the interview transcripts. The initial coding was done through the descriptive-level analysis. Once the transcripts were described and coded, the data was sorted and structured in a conceptual analysis that revealed the following data preservation information categories: knowledge, practice, trust, and experience. The researcher used the information derived from these interviews as the basis for a model and the survey.

Face-to-Face Surveys – Detailed Fact Finding

The researcher created a Qualtrics survey instrument (Appendix B) administering it to twenty-seven tenured or tenure track faculty in thirty-minute face-to-face sessions at each subject's location. The face-to-face survey approach provided additional verbal and non-verbal information. For example, as interviewees affirmed that they did use "...file naming convention for your research - Follows standards established by your discipline", their facial expressions led me to ask a follow-up. Over 90% confessed that they did not know if their discipline had a standard for them to follow. Survey results are presented using descriptive statistics in the

analysis. The survey data was downloaded into a spreadsheet and descriptive statistics were created in Excel to develop findings and recommendations.

Final Interviews –Issues Requiring More Elaboration

In the final stage, the researcher conducted interviews with two high-level campus leaders. As this study has shown, there are multiple layers of responsibility for digital preservation. The final interviews were to provide insight into what can and should be done at the highest levels of the institution which could facilitate digital preservation in the organization. Strategic goals and decisions guide institutional resource priorities. The resource priorities are translated into the hardware, software, and people for these services. The researcher asked open-ended questions (Appendix D) to explicate their viewpoint toward digital research files. Field notes were taken and analyzed to develop findings and recommendations.

CHAPTER 4 RESULTS AND ANALYSIS

The researcher used data from the institution, interviews, and surveys to elaborate on each the model's five elements: subject, instrument, community, rules, and division of labor. The analysis of each part of the model builds an understanding of the environment for research data preservation. Understanding the environment using the model can pinpoint specific areas that are problematic leading to better recommendations that prompt improvement to data preservation. The five model elements and supporting data are explained in this chapter.

Subject

CSU is the Colorado's land grant university with a mission of research, education, and outreach to the state population. In 2013, there were 30,647 students distributed across eight colleges conferring 72 undergraduate, 77 masters, and 45 doctoral degrees. CSU Extension serves 60 of 64 Colorado counties. The Agriculture Experiment Station conducts site-specific research in seven research centers distributed around the state. CSU is a Carnegie 1 Research University (Institutional Research - Colorado State University 2012, p. 5). It received \$259,017,009.67 in total research dollars in 2014 (Colorado State University Vice President for Research, 2015). The campus hosts a digital repository as part of an effort to manage data, promote research, and comply with new and evolving regulations.

The purpose of the Digital Repository is to promote and make accessible the intellectual output of the University to local, national, and international communities. This will maximize impact for individual CSU researchers and highlight the research profile of the University (Morgan Library - Colorado State University, 2012).

The research studied faculty in three academic departments at Colorado State University: Anthropology (Anthro) in the College of Liberal Arts, Atmospheric Sciences (Atmos) in the College of Engineering, and Computer Information Systems (CIS) in the College of Business.

Overall, differences between the three departments offer excellent contrasts in mission, funding, and academics. These contrasts provided rich foundational elements on which to build the analysis within and among each department, as well as their relation with the university and outside world.

The research dollars awarded to Atmospheric Science is high; resident student majors are large in Anthropology, while the CIS program has seen significant growth in student numbers, but has relied on other sources of funding such as its popular online business program. The tables 6, 7, and 8 provide some comparative information between the departments:

Table 3 Number Of Tenure Track Faculty By Department (Institutional Research - Colorado State University, 2015).

	Anthro	Atmos	CIS
Faculty Count	12	18	12

Table 4 Award Dollars Received From External Sources By Department x 1,000 (Colorado State University Vice President for Research, 2015).

	2008	2009	2010	2011	2012	2013	2014
Anthro	\$243	\$14	\$440	\$500	\$546	\$105	\$308
Atmos	\$16,240	\$19,456	\$16,880	\$15,958	\$14,040	\$12,937	\$14,527
CIS	\$170	\$141	\$30	\$13	\$231	\$171	\$208

Table 5 Student Enrollment By Department (Institutional Research - Colorado State University, 2015).

	2007	2008	2009	2010	2011	2012	2013	2014
Atmos	94	86	83	88	93	87	86	82
CIS	64	74	100	163	148	173	188	249
Anthro	224	236	231	272	316	334	320	271

Atmospheric Science is part of the College of Engineering and is on the foothills campus, about two miles from main campus. It does not have an undergraduate program. Its graduate

students work with faculty in labs funded by research. Researchers study global climate patterns locally and globally. Its hurricane forecasts to the public have been notable for decades. Recent research includes global warming modelling. Researchers may use their own data or massive data sets (terabytes and petabytes) from NASA, NOAA, or NCAR for analysis. The Federal government funds the preponderance of their research. This research is sensitive to federal policies on data access and preservation.

CIS is part of the highly centralized College of Business. Its faculties teach students to design information systems for organizational decision making. The CIS department participates in the prominent College of Business online M.B.A. program. The online program is hosted within the college's information technology infrastructure. There is an immediate financial incentive to preserve the digital instructional content for the college overall. Within the CIS program, students are taught data mining and data analysis, which are extremely important in understanding trends for business. While there may be a short-term focus on goals, there is also an awareness of long-term data needs. CIS largely receives funding from student enrollment, some research grants, and gifts-in-kind, such as state-of-the-art software from private industry for instruction.

The Anthropology department is part of the largest college at CSU, College of Liberal Arts, which is academically diverse, with a mix of undergraduate and graduate students, both Masters and PhD. The programs include Geography, Archeology, Biological Anthropology, and Cultural Anthropology. Anthropologists find, document, and archive historical artifacts as a core part of their practice. Anthropologists have a predisposition to preservation and archiving that should inform their digital preservation behavior. One participant described how anthropologists can store their collections permanently in a storage box at a specialized facility for a one-time

charge cost. There is an appreciation for preservation and how policy, practice, and cost should be a part of the long-term storage in this discipline. Their data include maps, interviews, recordings, and actual physical artifacts. Although there have been major efforts to digitize the artifacts as a means to access for research, they place greatest value on seeing the object in its original state. Their research can require sophisticated software for geographic information systems (GIS) or genetic modelling. Some data sets may be very large, but typical digital storage requirements are generally less than 200 Gigabytes per person.

Interviews revealed the researchers' different data needs as it applies to their disciplines. Anthropologists may use significant amounts of primary source data that they exhaustively gather *in situ*. They have intimate knowledge of minute details since they have such a personal link with its collection. The data includes artifacts, documents, reports, databases, and GIS files that are used to interpret populations. CIS faculty generally manipulate data from secondary sources to explain and improve processes. The secondary sources tend to be from the commercial or private sector. While most databases are small to medium in size, there is interest in business analytics which require very large data sets to run optimally. Atmospheric Science researchers download huge data sets from federal and international sources to run many different climate models. Like CIS faculty, their greatest need is to preserve the modelling programs and algorithms they use to run the models, not the data sets which are already in the public space. Their scientific experiments must be repeatable, and their results are under particular scrutiny given their role in establishing risk for insurance companies or providing information for the highly politicized climate change discussions. Because the federal government is a primary funding source and the sensitivity of their results, they had the highest awareness for the state of their data among the three departments during the interviews.

Instrument

The researcher combined interview and survey responses in the second phase to narrow the problem description, answer questions and begin to develop recommendations. 27 out of a total of 42 responded to the survey, relatively evenly spread across the departments. The response rate was as expected and factored in to the original design of the project, which uses multiple interviews, face-to-face survey approach, and institutional data to create a rich data set.

Table 6 Survey Response Rate

	Anthro	Atmos	CIS	Total
Respond	8	12	7	27
Possible	12	18	12	42
Percent	67%	67%	58%	64%

RIa – How much data do researchers have to manage?

Research data was separated into four categories: raw (newly created, generated or acquired), processed (reviewed, refined or revised), analyzed (critically examined), and finalized (changes to the data has ceased). The data are evenly spread across the three categories for Anthropologists and typically stored locally. With one exception, no researcher had more than 200 gigabytes in file storage in Anthropology and CIS. Atmospheric Sciences uses very large data sources (starting at one terabyte) that it runs models on. Hence, the finalized version is much smaller than the initial stage. There is some active data gathering through instruments and remote sensing, but many times, raw research data are downloaded from servers external to the institution such as NOAA and NCAR. They discussed their insights they gained by interacting with these repositories and how they adopted practices, such as metadata standards. Researchers in Atmospheric Science receive sufficient grant money to purchase their own IT infrastructure, with at least one system storing over one petabyte. CIS is also relatively evenly spread across

categories, but the preponderance of their raw data are secondary sources. Their data storage needs are equivalent to Anthropology, with one or two exceptions. Similar to Atmospheric Science, their long-term storage needs for programs and algorithms are actually less than what they need to run a project. The average responses for departments are listed in Table 10.

Table 7 Average Data State By Department

How much of your research data are...?			
	Anthro	Atmos	CIS
Raw	1.50	2.00	2.17
Processed	1.63	1.92	1.67
Analyzed	1.38	1.25	1.33
Finalized	1.00	0.92	1.33

1= 0 – 25%, 2= 26 – 50%, 3= 51 – 75%, 4= 76 – 100%

R1d – What digital file management practices do researchers use?

Table 11 is the researchers’ self-assessment of their knowledge in four file management tasks

Table 8 File Management: Knowledge

How would you describe your knowledge of the following?			
	Anthro	Atmos	CIS
Good file naming conventions	2.88	3.75	4.14
Meta-data	2.88	3.33	4.29
Back-up strategies	3.63	3.08	3.57
Long-term data preservation	3.25	2.67	3.29

Scale: 1=low; 5=High

When the study asked researchers what they did in practice, the data shows that they were particularly confident about the names which represented file contents. Interestingly, when asked about following discipline standards most respondents struggled. In a follow-up question, 15 of the 27 were uncertain or doubted that there was a standard, but chose to answer the question feeling that their standard approximates an informal standard of their discipline. Table 8 represents participant self-assessment of file naming. Most try to represent the contents of the

file, but are weaker in areas that would help manage data throughout the life of a project. File consistency and version control would be primary topics for a training program.

Table 9 Good File Naming Habits

How often do you use the following as part of your standard file naming convention for your research?			
	Anthro	Atmos	CIS
Represents file contents	4.38	4.75	4.43
Naturally ordered numeric or alphabetic	3.50	3.83	4.14
Consistent throughout your research project file system	3.50	3.50	3.29
Follows standards established by your discipline	2.13	2.67	2.00
Facilitates version control	3.50	3.75	3.86

Scale: 1=low; 5=High

Division of Labor

R1b – What storage resources do researchers use?

60 percent or more research data are maintained by the researcher or someone they help to pay directly in the departments. Atmos and CIS store about a quarter of their research data on college servers. Anthro is the greatest consumer of third-party storage, none of it university licensed. Almost no research data are stored and maintained by university IT staff. Files on third-party storage are used and shared easily from off-campus. The average of the responses is listed in Table 13.

Table 10 Data Responsibility Average Response By Department

What percentage of the storage media for your research data are...?			
	Anthro	Atmos	CIS
Managed by you personally	56%	18%	49%
Managed by someone within a research / collaboration group	18%	42%	20%
Managed by College IT staff	13%	27%	23%
Managed by University IT staff	0%	1%	1%
Third Party	13%	4%	8%

During the initial interview, trust was one factor that influenced the low use of college or university resources. Given this new knowledge after the first interview, the researcher sought a way to measure trust within the institution between researchers and IT. The Organizational Trust Inventory (OTI) (Kramer & Tyler, 1996) was used as part of the survey to evaluate researcher's trust of college and university organizational IT staffs. OTI has been cited in organizational and management literature. Cummings and Bromley assert that "trust reduces transactional costs in and between organizations" (Cummings, 1996, p. 303). They defined trust "as an individual's belief or common belief among a group of individuals that another individual or group makes (a) good-faith efforts to behave in accordance with any commitments both explicit or implicit" (Cummings, 1996, p. 303). Overall, one would expect means in the high six range for researchers to trust IT staff with their data. Figures 8 – 10 indicate that there is not overwhelming trust in organizational IT staff. Anthro and Atmos trust university IT staff more, however, as is shown in Table 13, university IT resources are virtually completely unused.

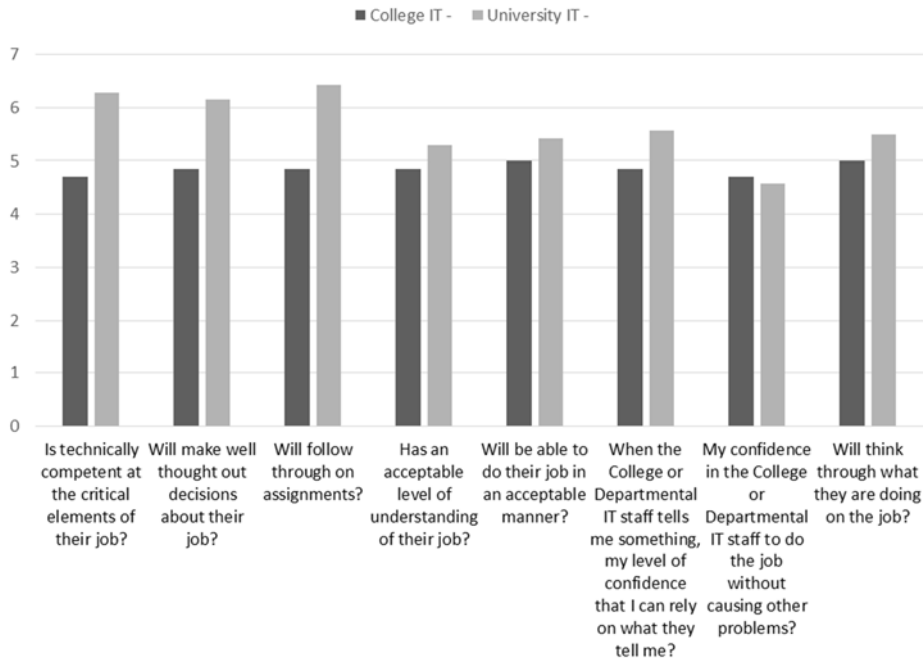


Figure 8 Organizational Trust Inventory – Anthro

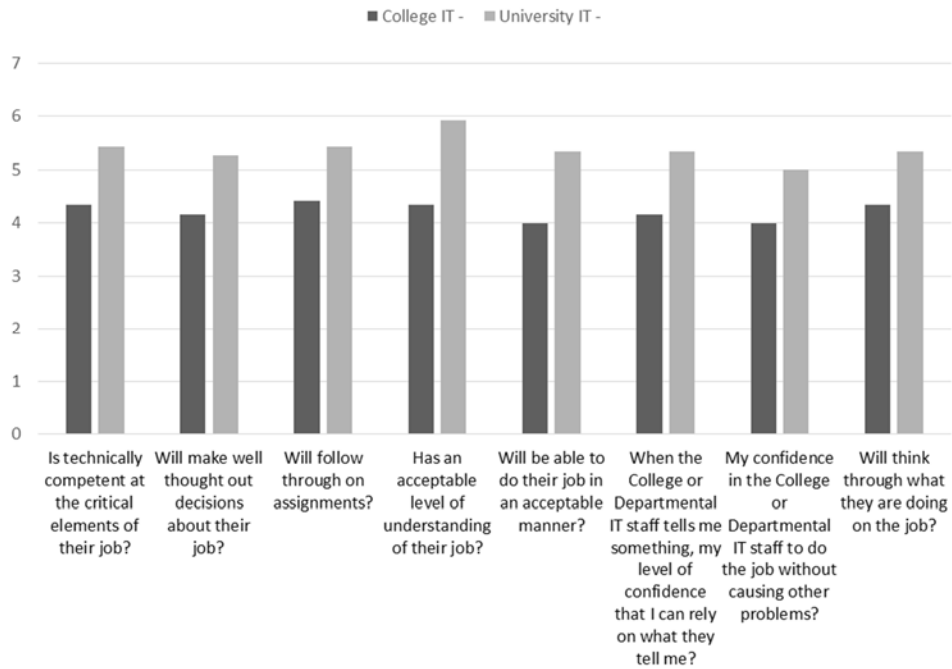


Figure 9 Organizational Trust Inventory – Atmos

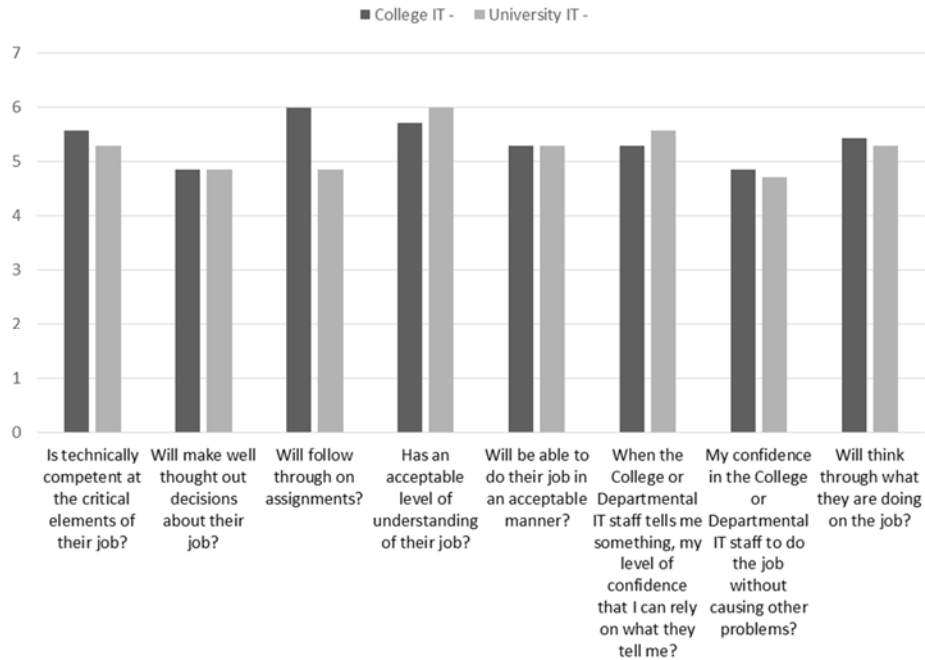


Figure 10 Organizational Trust Inventory – CIS

R1f – What critical research data loss events have occurred?

There is a strong incentive to understand how data are backed up. All researchers had at least one incident that they lost data from a research project. The responses for all departments aggregated are summarized in figure 11. Over one third of the individuals of those interviewed took more than one month to recover their work. One person stated that they never could recover their data. Interviewees used various backup strategies which presents a diverse risk portfolio to manage for data recovery. Some store data on flash drives, some on co-located local external drives, while some bring the drives home on a regular basis, and others use server space that their research project pays for.

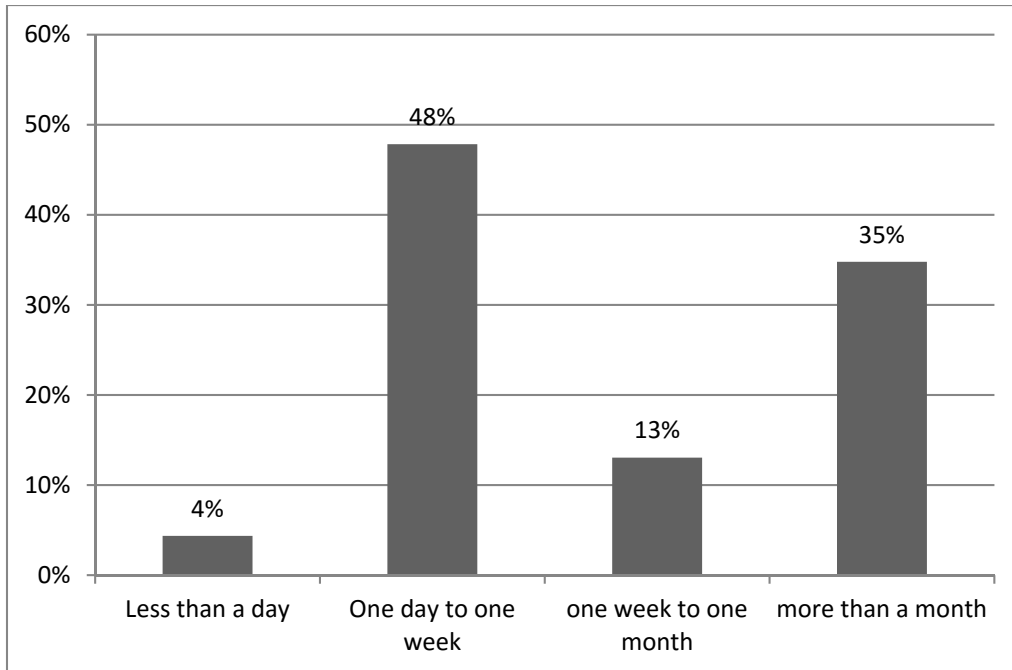


Figure 11 The most time a data incident in which critical research/data was lost cost

Most researchers claim to conduct an automatic data back-up daily; however during the follow up, many were not clear how it was taking place. There was an assumption that the IT department is doing it for them. The assumption may be correct, however, most are not using their departmental server space that is backed up. IT departments do not typically back up desktop computer hard drives. Furthermore, the research could not verify any IT service catalog that clearly defines their backup and recovery role and responsibilities to their clients. The activity may be taking place, but it is not codified describing when and how it is being done.

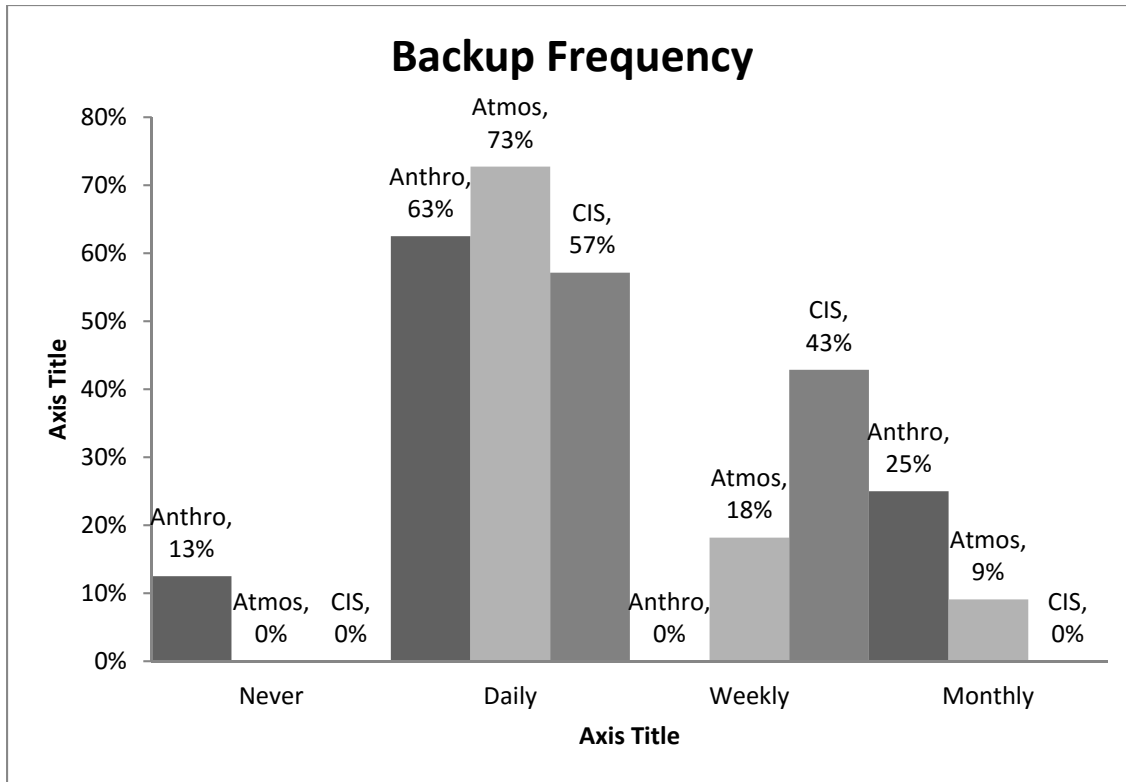


Figure 12 Backup Frequency

Table 11 Manual Or Automatic Backups

	Manual	Automatic
Anthro	38%	63%
Atmos	18%	82%
CIS	29%	71%

Community

When asked how interviewees acquire knowledge, the research found that it is atypical for a researcher to contact professional staff on campus to help maintain their data. Instead, they talk with their colleagues, solicit input from students in their lab or simply to try figure out their own best practices. Some of this may be due to a researcher's disposition, but as seen in the previous section, confidence in College or University IT staff to manage their files was also an issue. It is easier for researchers to discount the value of and ignore information from IT whose

service offerings are confusing or not clearly communicated. As elaborated in this study, digital preservation is a complex issue for which researchers have zero tolerance for failure. Any reported breach in this trust affects the entire community. This can create a barrier in an already difficult communications environment that is difficult for professional staff to break through. This is a problem for the institution. For example, the university has sent out information about contract agreements with both Microsoft and Google that provide basically unlimited cloud storage for all university employees. Most faculty interviewed were unaware of the resource.

The group generally felt that they needed a better process to preserve their research data, but that they were getting by. There are local activities that are designed to help researchers understand the issues and preserve their data. Several workshops have been offered on campus and there are knowledgeable professionals available on call. As stated earlier, researchers do not seek the advice of professional staff. When asked about formal training, very few had attended.

R1c – What digital data management training activities do researchers attend?

No one answered affirmatively when asked “How many data management workshops or sessions have you / attended to the best of your recollection – CSU hosted?” One person answered affirmatively when asked “How many data management workshops or sessions have you / attended to the best of your recollection – non-CSU hosted?” Only 22 percent said that they had had some form of digital data management training in their lives. Thus, researchers confirmed that they use their own best practices with minimal training to preserve data.

In general, researchers intend to maintain the data within the time frame of their research projects. The perceived planning time horizon for the majority was three to five years for their digital data, although there are some research faculty with much longer research projects. Most acknowledge that this time frame may be a problem, but that others bear responsibility after the

research is done. Many of the interviewees struggled to describe in words their preservation process, the responsibility for their research data's long-term preservation, or when and how they should allow access to it as required by NIH and NSF. Interviewees generally believe that their sole preservation responsibility is for the data until the final project report submission, whether it is a publication, poster, or other document. Once the report is submitted, they do not consider that the data used is of any value to anyone else, given the risk of taking it out of context. Half of the interviewees said that their files had metadata descriptors, but very few could say what descriptors were included or how it was included. This does not mean that metadata isn't included, but it is a sign of a knowledge gap in a critical area for preservation activity.

The data shows that researchers need and want to make their data accessible to collaborators outside the institution. Faculty use cloud services for ease of access for themselves and their collaborators. The potential risks of cloud storage services are overlooked for the convenience and ease of use of third-party solutions. The risks include unclear legal precedents for the space, business risk, overseas server locations, and liability for damage after a cyberattack. Although the questions wasn't raised, it is extremely doubtful that anyone has read the latest End User License Agreement for the service they are using. It is equally doubtful that the legal language would be understood by all but a small fraction of those who have. The university has mitigated these risks with licenses for both Google and Microsoft storage to anyone with an eID.

Microsoft's OneDrive offers one terabyte (soon to be unlimited) of data storage that is integrated with their Office product to university faculty and staff. Google provides basically unlimited storage to students, staff, and faculty (once a student graduates, they can easily convert their account to an alumni account and they can take the data with them upon graduation). Even

though the pace of change in the physical media (floppy disk, hard drive, flash memory) has slowed and evolved to the cloud, the cloud environment itself is becoming much more complex as providers try to meet our needs. There are multiple vendors with cloud storage solutions (box.com, Dropbox, Amazon, and IBM). Data stored in unregulated commercial space represents a new risk. Dropbox was mentioned several times as a preferred location to place data sets to share both internally and externally. The university has no agreement with them to assure control and preservation of the data are maintained.

R1e – What data do researchers expect to share from their research?

In general, research faculty expect to share their data at some level. Figure 11 shows that Anthropology faculty are most sensitive to releasing their raw data during research. All three groups are willing to trust all data with trusted colleagues. Atmospheric Science researchers are willing to trust processed and analyzed data with anyone, on request, during research.

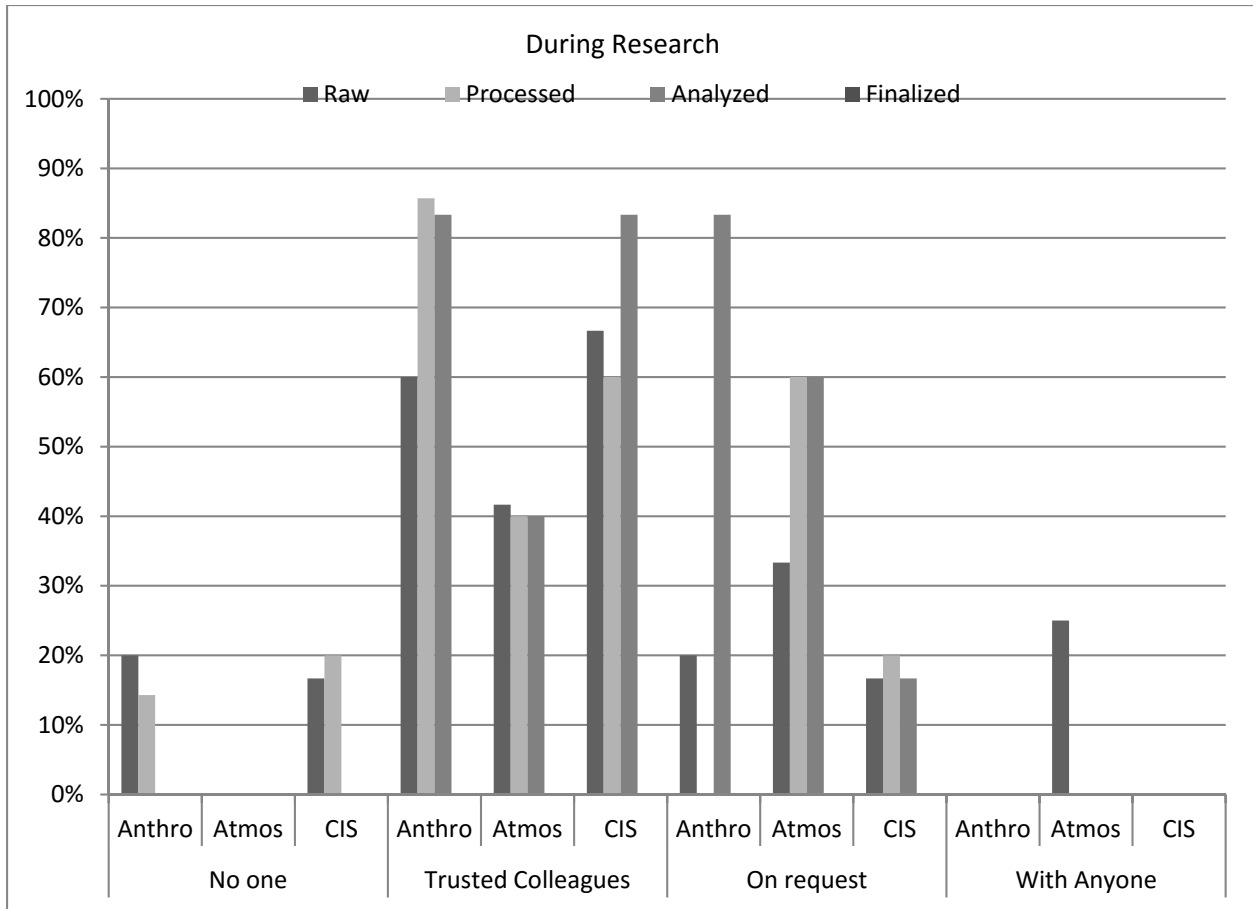


Figure 13 What data do researchers expect to share from their research? (During research)

Once a research project is complete, researchers claim that they are quite open to sharing, as seen in figure 13. The willingness to share can be seen as a positive when considering next steps to complying with federal open access requirements.

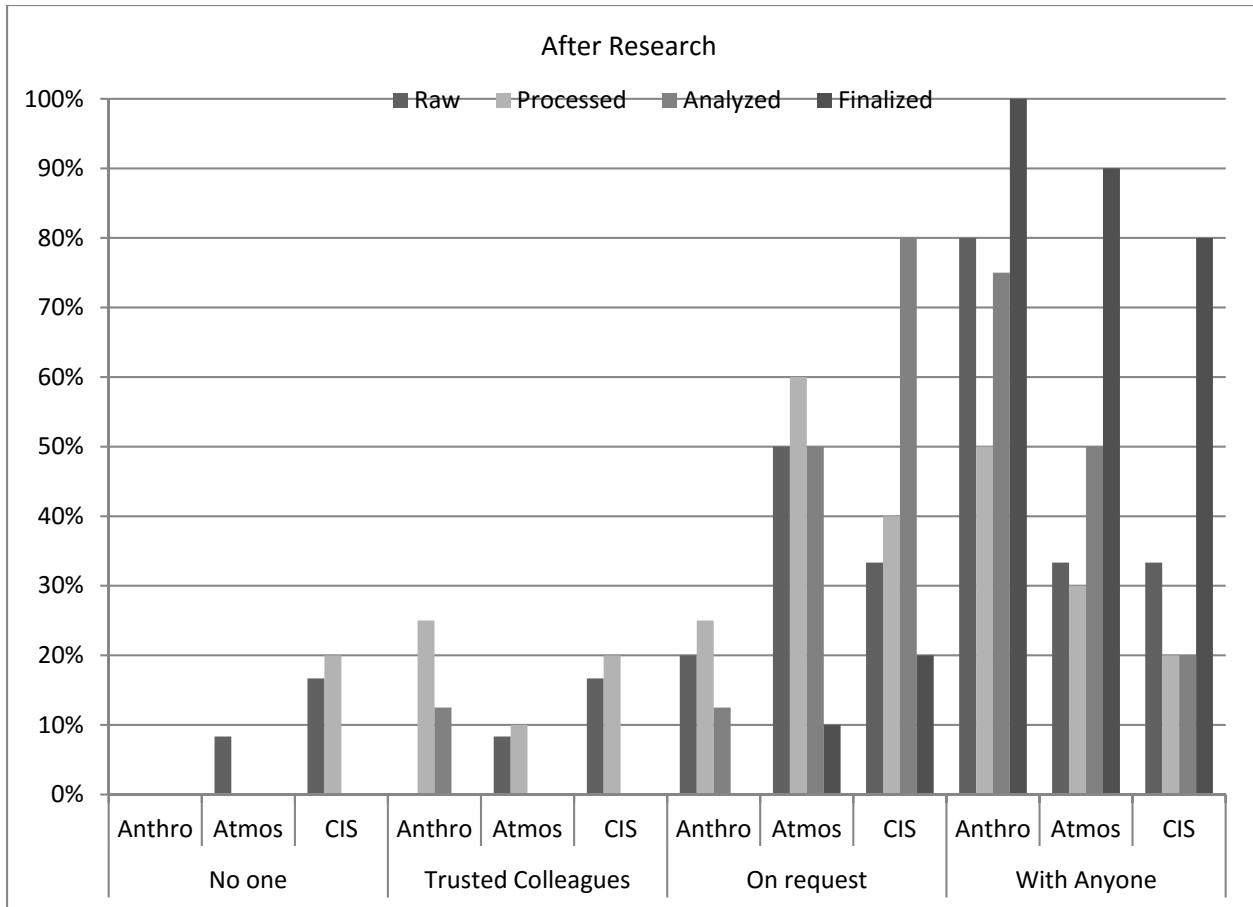


Figure 14 What data do researchers expect to share from their research? (After research)

Rules

R1g – To what extent have researchers needed to file data management plans?

Table 14 lists, by department, the percentage of each department that have been required submit a data plan and use a data repository. It also lists their level of compliance and willingness to do so in the future. Atmos generates the greatest numbers of NSF grants. All departments rely on repositories external to CSU heavily to fulfill their obligation. Metadata provides the context with which the data was collected. Overall, metadata is not archived as well as the source data, even though it is a condition of a DMP. Atmos and Anthro have deposited some metadata to other repositories while CIS has not. No department has deposited metadata at

CSU. Tables 13 and 14 contain a listing of where the files or metadata was stored by department.

Table 12 Data Preservation Requirements

	Anthro	Atmos	CIS
Do any of your funding sources require that you create a data management plan as a condition of funding?	38%	92%	43%
Do any of your funding sources require that you share your data with others, publish your data, or deposit your data into a data repository?	25%	92%	29%
Do any of your funding sources require that you preserve your data beyond the life of the funding?	63%	67%	29%
Do any of your funding agencies require that you place your research data into a data repository (a central place where data is stored and maintained)?	13%	67%	14%
Have you ever deposited data into a data repository (a data repository refers to a central place where data is stored and maintained)? CSU	0%	8%	0%
Have you ever deposited data into a data repository (a data repository refers to a central place where data is stored and maintained)? Non-CSU	25%	58%	14%
Have you ever deposited any metadata into a data repository? CSU	0%	0%	0%
Have you ever deposited any metadata into a data repository? Non-CSU	13%	33%	0%
If not required, would you be willing to submit your data to a data repository in the future?	88%	92%	71%
If not required, would you be willing to submit your meta-data into a centralized repository in the future?	75%	75%	86%

Table 13 Data Repository Locations

Repository Location	Dept
Paleo databases of Americas	Anthro
The Digital Archaeological Record (tDAR)	Anthro
Internal Group	Atmos
Lawrence Berkeley National Lab	Atmos
NASA field projects	Atmos
NASA - unspecified	Atmos
National Center for Atmospheric Research High Performance Storage System (NCAR-HPSS)	Atmos
NCAR - unspecified	Atmos
National Oceanic and Atmospheric Administration (NOAA) - unspecified	Atmos
Sponsor - unspecified	Atmos

Table 14 Meta-Data Repository Location

Repository Location	Dept
tDAR	Anthro
NCAR - unspecified	Atmos
Lawrence Berkeley National Lab	Atmos
NASA Oakridge	Atmos
NASA - unspecified	Atmos
NASA - unspecified	Atmos

Research faculty show a strong willingness to submit both data and metadata to a repository in the future. The researcher extracted Table 12 data into Table 15 for easier analysis. At face value, there appears to be a positive response to the data management plan requirement. This fact is further reinforced by the willingness to comply with the mandate, as shown in Table 15.

Table 15 Data Management Plan And Researcher Action

	Anthro	Atmos	CIS
Do any of your funding sources require that you create a data management plan as a condition of funding?	38%	92%	43%
Have you ever deposited data into a data repository (a data repository refers to a central place where data is stored and maintained)? CSU	0%	8%	0%
Have you ever deposited data into a data repository (a data repository refers to a central place where data is stored and maintained)? Non-CSU	25%	58%	14%
Have you ever deposited any metadata into a data repository? CSU	0%	0%	0%
Have you ever deposited any metadata into a data repository? Non-CSU	13%	33%	0%

R2 – How can digital preservation be communicated to researchers to improve the permanency of their data?

Based on earlier presented data, research faculty seek trusted sources (colleagues, their own students) for information. The researcher used Social Network Analysis to find a way to connect individual researchers at CSU through the funding source. The researcher selected the three departments based on their diverse programs and funding. Figure 15 and the companion Table 15 is a visualization of all federally funded grants from 1987 – 2014 for the three departments in the study. Anthropology in the upper left, CIS in the lower left and Atmospheric Sciences on the right. Atmospheric Sciences has 42 unique connections with 1821 awarded

grants to federal agencies. Most of their awarded grants, indicated by the heavier line, are with NASA and NSF. It is best to develop preservation training for shared connections since it optimizes resource use and potentially creates synergy between the departments. Atmos shares eight connections solely with Anthropology and four connections solely with CIS. Anthropology and CIS share a single connection with USDA-USFS Rocky Mountain Research Station. All nodes share connections with both NSF and USDA-USFS Forest Research.

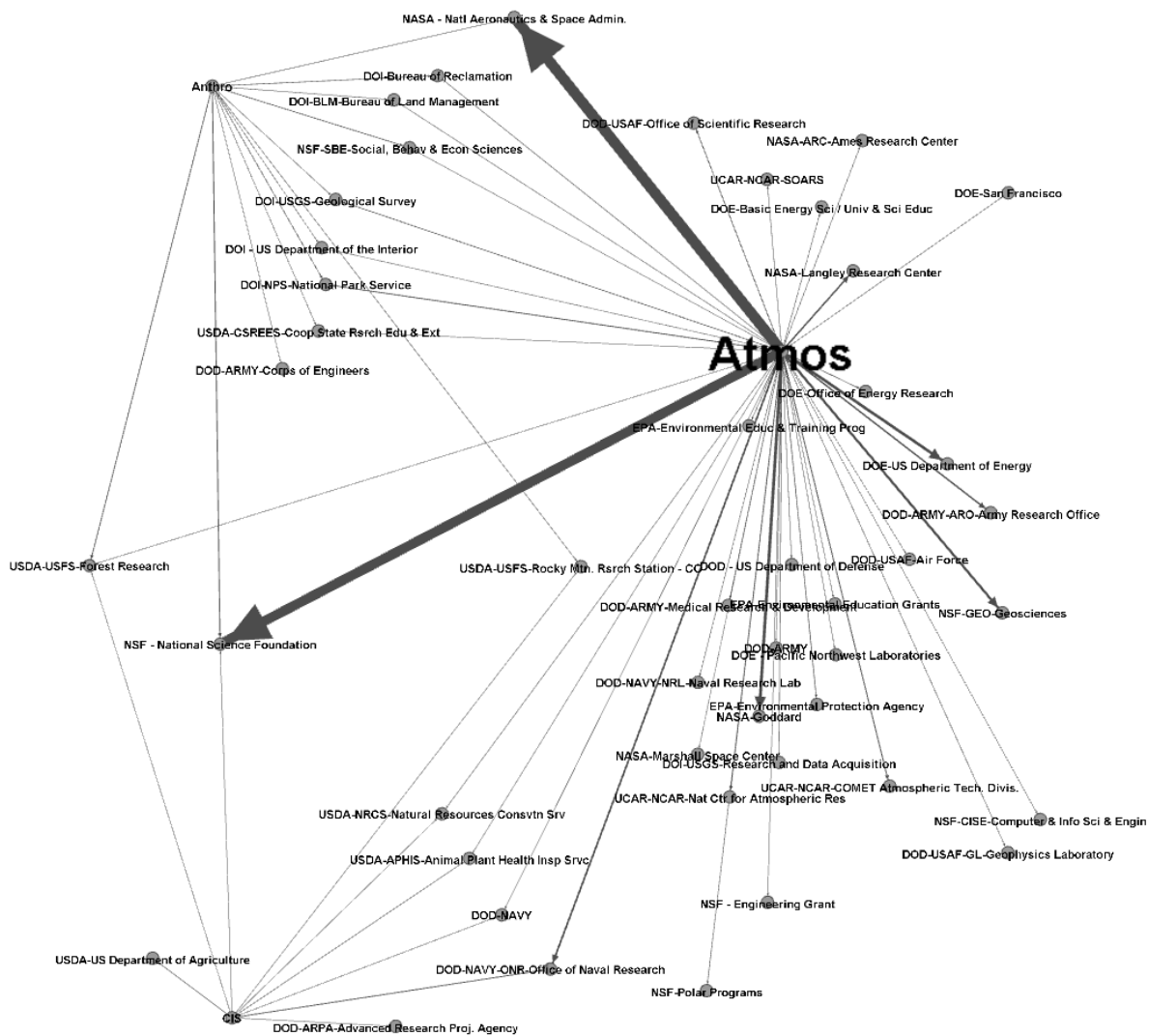


Figure 15 Weighted Out-Degree for Federal Grants 1987 - 2014 - Target Group

Table 16 Weighted Out-Degree: All Federal Research Grants 1987 - 2014 By Study Group

Id	Label	Out-Degree	Weighted Out-Degree
Atmos	Atmospheric Science	42	1821
Anthro	Anthropology	12	94
CIS	CIS	9	44

The previous visualization and companion table show the three departments are connected through various federal granting agencies. However, only the NSF has a data management plan requirement presently. The visualization in figure 16 and the companion table 17 isolate on NSF linkages from the three departments to NSF, 1987 - 2014. Notice that someone in each of the three departments have received an NSF grant during the period. The shared experience of receiving an NSF grant can be lead to common areas for training with the individuals in each of the departments. The training can provide an entry point for college and university support staff to improve communications with the research community.

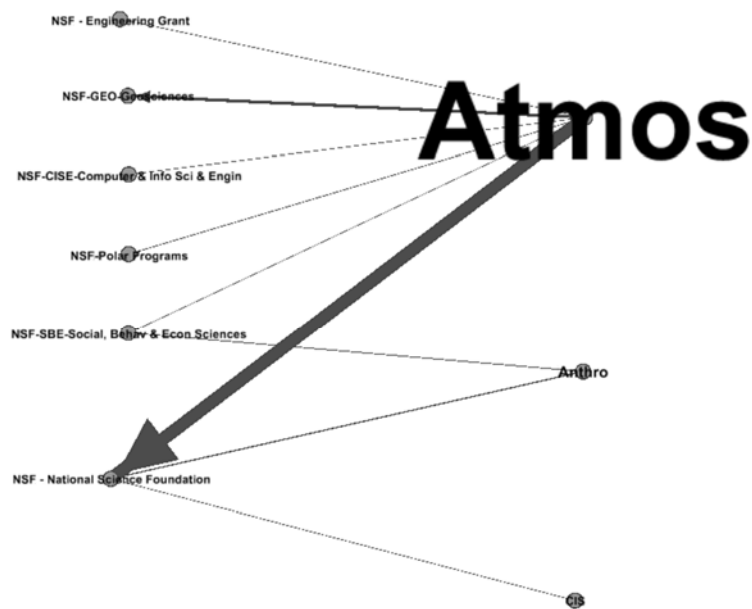


Figure 16 Weighted Out-degree for NSF University Grants 1987 - 2014 - Study Group

Table 17 Weighted Out-Degree NSF Research Grants 1987 - 2014 By Study Group

ID	Label	Out-Degree	Weighted Out-Degree
Atmos	Atmospheric Science	6	622
Anthro	Anthropology	2	40
CIS	CIS	1	2

Eigenvector centrality plots throughout the data depict a power law distribution. Figure 17 is one example depicting federal grants from 1987 – 2014 which is the same data used for figure 16. In the diagram, notice that one individual has over 110 connections while over half have ten or fewer. Watts and Strogatz demonstrated that short path lengths and high clustering like this indicate a strongly connected network and the existence of a small world phenomenon (Watts, 2003). Using the world phenomenon and Grannoveter’s theory of weak ties, we know that it would be optimal to build a communications plan that focuses on a few key people from each department.

Eigenvector Centrality Report

Parameters:

Network Interpretation: directed
Number of iterations: 100
Sum change: 0.0

Results:

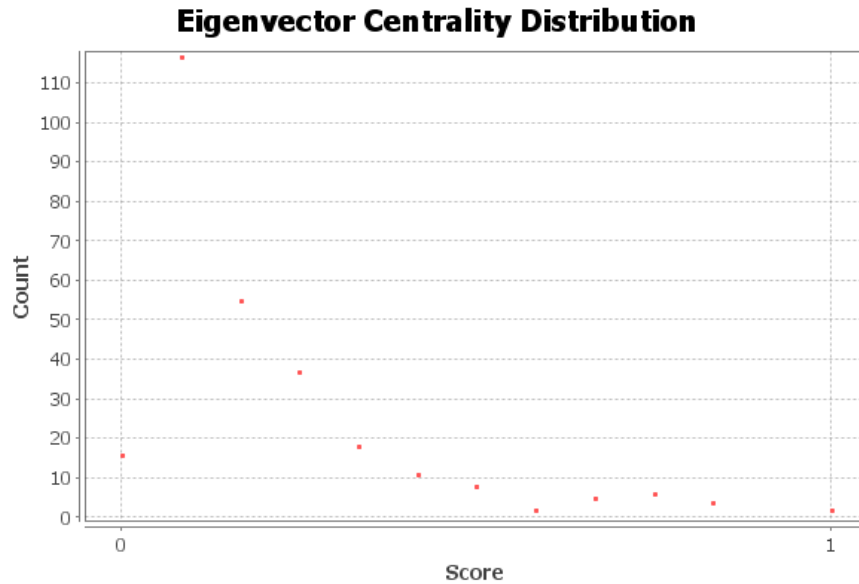


Figure 17 Eigenvector Centrality Distribution – Federal grants 1987 - 2014

The researcher then plotted all grants for researchers in all the three departments since the NSF established the data management plan requirement in 2011. Only Atmos has received a grant from NSF since the requirement went into effect. Figure 18 depicts the network of researchers and grants. The software (Gephi) was run to automatically adjust the layout based on the centrality of each node. Each granting agency name is listed. Researcher names have been anonymized with the number representing their department and a letter representing the individual. The number 1 represent CIS faculty, 2 Anthro and, 3 Atmos. The dots representing the researcher nodes have been lightened for better visibility. There is a clustering of “3’s” in the center of the figure indicates that they are closely connected via their grants. Outliers, like 2b and 1h to the right, are easy to see. They are not well connected to the rest of the campus

research community through their grants. The position of 2i at the top left and both 1d and 2e toward the bottom left in the graph are of interest. These represent nodes that are connected to a granting agency mutual to Atmos.

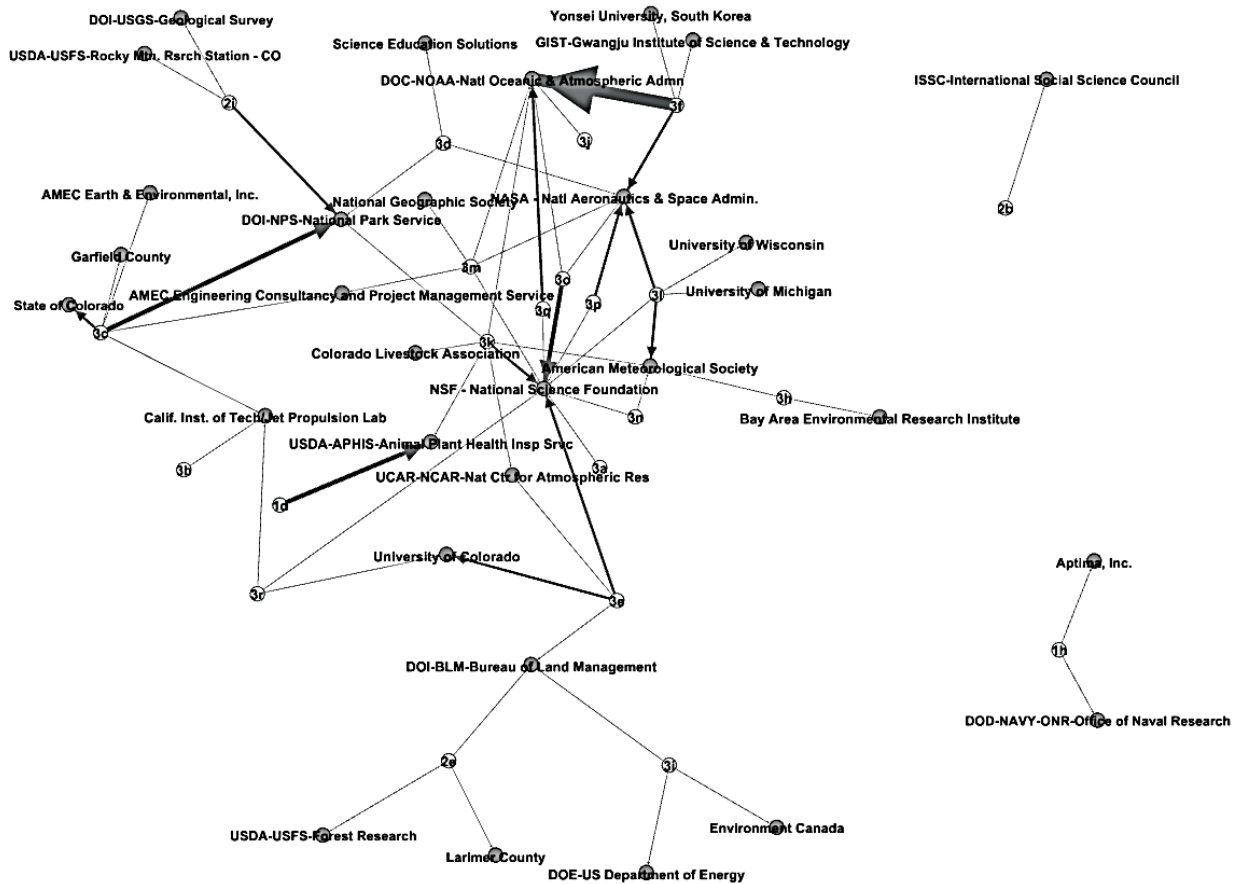


Figure 18 Successful Grant Submissions for three Departments since 2011

The researcher focused on how connections between researchers with NSF funding grants and researchers connected in the three departments (Table 18). The first column shows, based on the survey data that eighty five percent of those who have received an NSF grant since 2011 are not depositing meta-data. Seventy five percent are not using a data repository. There is a significant non-compliance issue, based on the data. Interestingly, the results are not much different for those without an NSF grant, as shown in the second column. The next three columns break down the percentage of non-NSF funded research by department. The Atmos

faculty who have not received NSF funding, but are in a department that is much more reliant on it, are slightly more compliant than the group that has received NSF funding. The CIS, the department least dependent on any federal funds, is also least compliant. Existence or non-existence of a DMP requirement may influence researcher behaviors. It may simply be too early in the research cycle to be shown in the statistics.

Table 18 Comparative NSF Influence On Data Deposit

	All		Anthro	Atmos	CIS
	NSF Yes	NSF No	NSF No	NSF No	NSF No
No Meta-Data	85%	94%	93%	83%	100%
Yes Meta-Data	15%	6%	7%	17%	0%
No Data Repository	75%	85%	86%	67%	93%
Yes Data Repository	25%	15%	14%	33%	7%
Total N	20	34	14	6	14

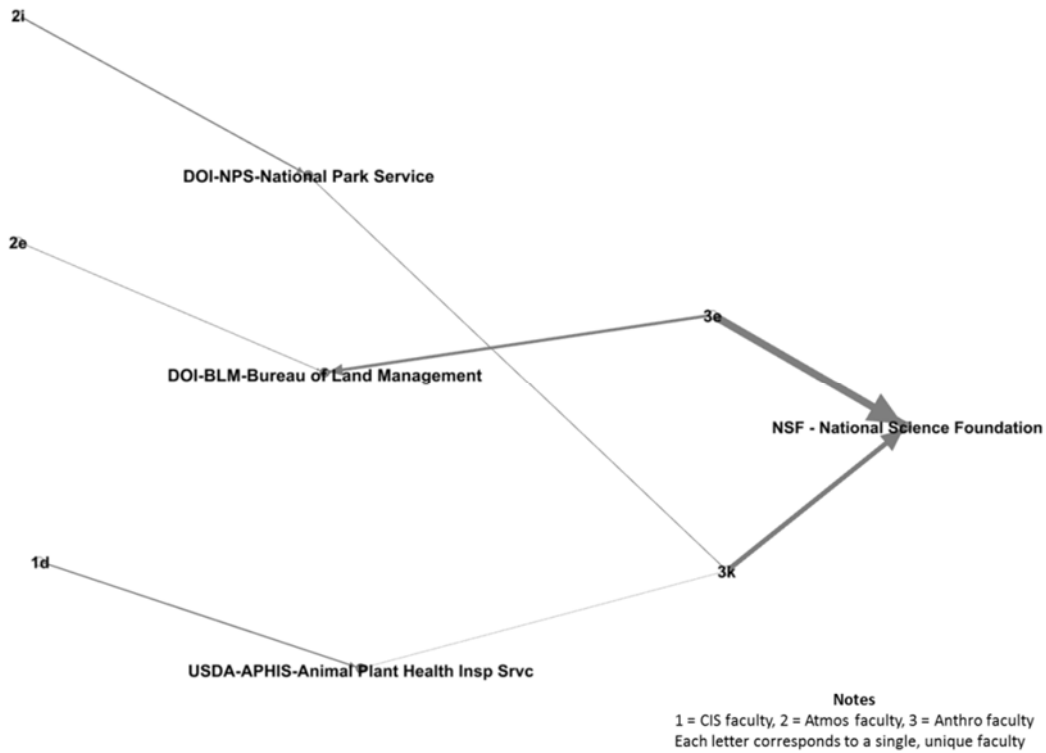


Figure 19 Indirect connections to NSF

Figure 19 depicts the only two researchers (3e and 3k) who have received NSF funding since 2011 who are also connected to faculty in one of the other study departments (2i, 2e, and 1d) through another funding agency (DOI-NPS, DOI-BLM, and USDA-APHIS). The group of five researchers provides a network hub to stimulate dialogue with each of the three departments about digital preservation best practices. Given the highly connected network, establishing insider status by the professional staff within each department can improve the transmission of best data management practices to each.

CHAPTER 5 DISCUSSION

Contextualizing Data Preservation

As discussed in the background chapter, the study combined the OAIS reference model categories with Engeström's enhancement to the Activity Theory model as the design construct and elaborated on each of the five elements: subject, instrument, division of labor, community and rules. This design is an expression of OAIS innovation adoption processes.

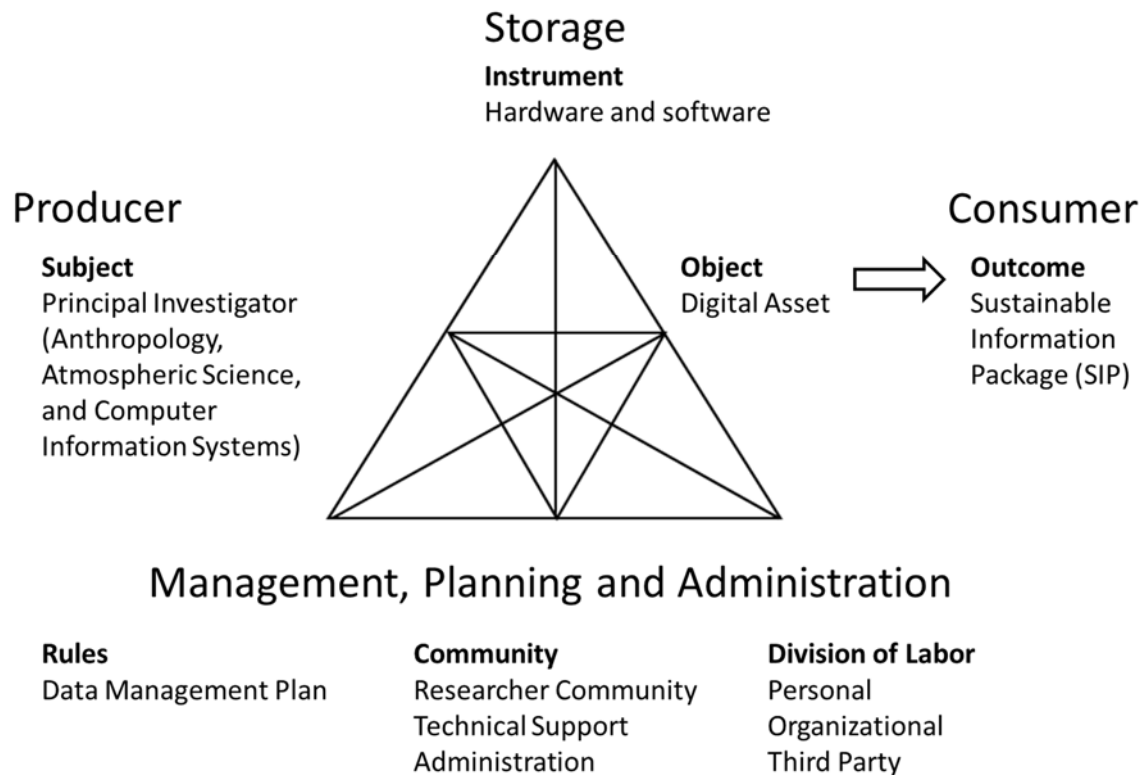


Figure 20 Study Model Using Activity Theory

Every effort was made to take the perspective of the researcher since, as discussed earlier, they are the creator of each artifact who should have the first opportunity to establish the value of each object and find an appropriately place for its disposition. The researcher (subject) creates a file (digital asset) that needs to be converted to a sustainable information package (SIP). The

researcher developed specific concepts for analysis through the interview and survey phase and categorized them in the model.

Subject

Researchers have the greatest responsibility for the disposition of their data. Their initial decisions to save, describe and place data properly can have the greatest impact on its lifespan. The research studied three academic departments at Colorado State University: Anthropology (Anthro) in the College of Liberal Arts, Atmospheric Sciences (Atmos) in the College of Engineering, and Computer Information Systems (CIS) in the College of Business. The researcher chose the departments because of the different perspectives they have toward research and the preservation of its data. Diverse interaction with the instrument, rules, community, and division of labor can inform this study's recommendations.

Instrument

Faculty have several choices for physical data storage. Choosing to store data locally is convenient and simple. It fits small to medium data sets during a research project cycle. It is easy to transport via the network or portable storage devices. It is typically neither sharable nor redundant. Version control issues are also inherent as files move between devices. Departmental, College or University data storage is redundant with controlled access mechanisms in place. Faculty can access any file from multiple locations eliminating the intrinsic version control issue. In most instances, IT staff can roll back files to earlier versions restoring lost work. Most organizations have funding for the service and equipment, but unless there is a cost recovery model, space is typically limited by quota size or purpose. The network speed will influence the use of this option. Files greater than one terabyte will take hours to upload. Access is still limited to those whose identity can be verified by the institution

electronically. Finally, organizations have seriously started to consider third-party storage solutions. Data storage can be scaled easily for multiple institutions achieving cost reductions for equipment and professional staff. The professionally managed storage can be provisioned to faculty cheaply. Under the existing Microsoft license, all Colorado State University faculty and staff can store one terabyte on Microsoft's cloud for free. The network speed will influence the use of this option even more.

Division of Labor

Faculty are creative problem solvers who have the freedom to use resources within and outside the organization. They can find many ways to store, share and backup their data. Each solution has risk. Faculty assume most of the responsibility personally when they save files on their personal system. They need to ensure that their system is protected, updated from malware, backed up, preferably in geographically separated spaces and also make sure that all their data are transferred whenever they upgrade their computer. The number and size of data sets can also strongly influence this choice. As research projects grow, it is likely that they will seek other solutions. Many of these responsibilities can be delegated to professional staff when they use institutional data storage. IT staff will ensure that data are accessible, protected and backed-up. Faculty should manage file and folder organization. They also need to continue to verify file formatting and readability as new software versions are released. The research investigated cloud solutions use, both on premises and commercial, given their remarkable advancements in the past five years. Third-party cloud storage End User License Agreements should be read to confirm that FERPA, HIPPA or requirements to store data in the United States have been met. Institutionally sanctioned third-party cloud storage is one way to assure their appropriateness.

Community

As depicted in figure 18, the community in the study is represented by the faculty who conduct research, the IT professionals who support their research and the organizational climate the researcher works in. Communication barriers include within-discipline influences for faculty that are stronger than relationships with colleagues from other departments. Information is less readily received from local sources. Expectations of faculty for IT services can also be a barrier to a community digital preservation effort. If expectations are low, IT specialists will not be sought for information. The institution's administration has provided resources for personnel, hardware and software to solve some of the problems, but have not actively engaged in creating a narrative for the campus community to set goals. The health of the community with respect to digital preservation, per the AT model, is a determinant of institutional digital sustainability.

Rules

NSF data management plan requirement is a requirement that demands researchers consider the organization and disposition of their digital data. The data management plan prerequisite only applies to these researchers, but its affects can ripple through the institution. Additional organizations will likely adopt a data management requirement to preserve costly research data. NSF is an exemplar for government and private entities who want to implement data management plan requirements in the next few years.

The data matrix is essential for organizing social network data. "In variable analysis...each case (is)...represented by a row...while the columns refer to the variables" (Scott, 2000, p. 38). Social network analysis is a relational model without attribute variables. The relations between and among entities are defined and measured instead of variables. Each case is measured by its affiliations. The affiliations can be a common event, organization or

activity. Social network analysis uses these case-by-affiliation variables to develop incidence and adjacency matrices. “The cases are... the particular agents that form the units of analysis, but the affiliations are the organizations, events, or activities in which the agents are involved” (Scott, p. 39). The cases are the rows and the affiliations are the columns in a matrix. For example, data collected about four researchers on campus about their activities is placed on the incidence graph (Table 3).

Table 19 Sample Incidence Matrix

0=No 1=Yes	Digital Preservation attributes			
	Has submitted an NSF grant	Has attended an NSF grant workshop	Stores data in the Digital Repository	Discipline is a “Hard Science”
Faculty A	1	0	1	1
Faculty B	0	0	0	0
Faculty C	1	1	0	1
Faculty D	1	1	1	1

Faculty A has submitted an NSF grant and stores data in the university digital repository.

Faculty C has attended the workshop, but has not submitted a grant. The information by itself this may interesting, may have some utility, however additional analysis may be done through adjacency matrices.

Table 20 Adjacency Matrix: Digital Preservation Attributes By Digital Preservation Attributes

	Has submitted an NSF grant	Has attended an NSF grant workshop	Stores data in the Digital Repository	Discipline is a “Hard Science”
Has submitted an NSF grant	-	2	2	3
Has attended an NSF grant workshop	2	-	1	1
Stores data in the Digital Repository	2	1	-	2
Discipline is a “Hard Science”	3	1	2	-

The cells of an adjacency matrix contain the resultant values of binary values from the incidence matrix. Table 20 shows the results of a digital preservation attribute adjacency matrix and Table 21 shows the result of a faculty adjacency matrix.

Table 21 Adjacency Matrix: Faculty By Faculty

	Faculty A	Faculty B	Faculty C	Faculty D
Faculty A	-	0	2	3
Faculty B	0	-	0	0
Faculty C	2	0	-	3
Faculty D	3	0	3	-

These simple examples begin to reveal facts such as that the relationship between disciplines as a hard-science NSF grant submission or that there should be some affinities between faculty A or C and D. Social network analysis is a method to visualize and apply statistics to university connections. For example, Cointet measured and graphed the “semantic

landscapes of scientific knowledge communities”(Cointet, 2012, p. 2). He “defined a proximity metric...to map the scientific landscape made by the aggregation of publications over time” (Cointet, 2012, p. 3). He graphed the network linking key terms and phrases from these publications based on their semantic distance.

Two social network analysis statistics that can be used to help understand how data preservation may spread through the organization are Weighted Degree and Eigenvector centrality using institutional-awarded grants. Weighted Degree measures connectedness by summing the number of entities someone is connected to and is weighted by the number of connections it has with each. For example, a person who has a grant with five different agencies may be considered better connected to the network than someone with less than five. Alternatively, a person who has five grants with only one agency is less connected to the network overall, but has a much stronger connection to the one. These patterns can also be elaborated on using Rogers’ (Rogers, 2003) concepts of heterophily and homophily. Individuals in the first example are typically more receptive to change and choice. They are a path to the overall group since they are recognized, legitimate members. The relationships can also have directional qualities. A grant submission from the sponsor’s perspective would be considered “in-degree” and an awarded grant would be considered “out-degree”.

Graphs representing weighted out-degree network statistics can be found in Appendix A with colleges represented in a circular distributed pattern using Fruchterman Reingold layout. Each small dot represents a sponsor that has funded a grant for one of the colleges since 1987. The points gravitate toward the center of the graph as the number of shared connections increases. The agencies (dots) in the middle of the diagram are more highly connected to every college; therefore carry more influence than those on the periphery. Changes to proposal

requirements such as a data management plan, among agencies that are closer to the center should have greater impact on the university than those on the edges.

Eigenvector centrality measures proximity to the influential core of the network by measuring its connections as well as its connections' connections. An individual whose networks are not well connected has been proven to be located closer to the edge of a network. Conversely, an individual whose networks are also well connected is closer to the influential center of the network. Google's PageRank algorithm is a version of Eigenvector centrality. Brin and Page state that "PageRank or $PR(A)$ can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web" (Brin & Page, 1998). This method can be used to understand how sustainable digital practices may spread from external organizations, such as NSF, through an institution.

Finding Answers

The effort to document research can be significant. Transforming the way research is maintained from traditional, analog formats such as pictures, recordings, and lab notebooks to digital formats requires that each investigator rethink and recreate their records management methods and workflows. This effort competes with and is prioritized with the ongoing challenges of research, teaching and publication in very competitive fields. The federal government and its research grant system is stimulating the transition through its data management requirements. The NSF data management plan specifically requires "plans for archiving data, samples, and other research products, and for preservation of access to them" and that "*standards* to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies)" [italics added] (National Science Foundation, 2013). In 2013, the White House

published a Memorandum that mandates faster and more open access to the results of research grants (Office of the President, 2013). The White House Office of Science and Technology Policy issued also issued a mandate that nearly all Federal granting agencies require access and digital accountability as a condition of all awards (Office of Science and Technology Policy, 2013) following NIH and NSF data management plan requirements.

Lee (2005) focused on the creation and adoption of the OAIS model as the first step towards sustainability. The OAIS defines a global modality in which digital interaction translates information into knowledge and knowledge into information. Signification is exercised through language, categorization and the metadata created to sustain digital knowledge. But how do we implement its protocols into personal and professional practice? It is only one component of a growing digital life. Giddens' structuration theory holds that all human action is performed within the context of a pre-existing social structure governed by a set of norms (Giddens, 1984). He interprets structures as rules and resources. Similar to Giddens, Suchman's situated actions view the duality of our tools and the interaction between individuals and the implements they use to mediate their environment (Suchman, 1987). She realizes the problem of shared understanding or "mutual intelligibility" should account for the foundation of social order. However, her claim is that technology is a participant in the social order. "... We now have a technology that has brought with it the idea that rather than just using machines, we interact with them as well... the notion of "human-machine interaction" pervades both technical and popular discussion of computers, whether about their design or their use" (Suchman, 1987, p. 29). Activity theory maintains that we are socializing our tools as they change us. Humanity's transformation to a digital paradigm has diffused through societies since Turing wrote "On Computable Numbers" in 1937. Today, the remarkable systems we build grounded in

Turing's seminal ideas are becoming extensions of ourselves. Social network analysis is a relatively new way to measure the connected age we find ourselves in. It exploits the progress in computational power that it seeks to measure.

File management is at the core of digital preservation efforts. Digital preservation is an act based on human behavior whose scope transcends the technology that we use to navigate it. Digital preservation is based on decisions by the creators of the artifacts, the stewards who manage it, the technicians who maintain the systems and leadership that creates policy and funding for it. Researchers are creators of the information and, as stated earlier, have the greatest responsibility to establish the value of digital artifacts. Since the digital age will rely on the way we preserve our data, preservation practices must evolve to satisfy societal needs. Given the transient nature of the digital world, our decisions of what data to save has fundamentally changed. DOI is a mature theory, elements which Watts quantified using Network Analysis. The first research questions are designed to find opinion leaders and activity clusters with respect to digital preservation activity. Activity Theory provides a framework that includes rules, resources, tools and culture to answer the second. Both can explain the introduction and diffusion of digital preservation practices throughout the organization.

Recommendations

The study shows that researchers maintain data until it no longer suits their needs. Once they are done, only a very small portion of their data are in publications or reports that are preserved long-term. This leads to gaps in stewardship of the data. These gaps can create a nearly impossible task of recovering data. Therefore, this study proposes three temporal contexts to bound digital sustainability issues for the new paradigm – short-term, long-term and trans-generational. Gaps in data stewardship can then found and resolved as the three contexts are

linked improving overall data sustainability. Short-term storage is three to seven years. This is the time that many regulations require documents be stored for reporting purposes. Thus, the document owners are required to find solutions to satisfy these requirements. This definition also falls within the comfortable range of media and software life expectancy. Long-term storage is anything greater than short-term, but less than a generation. For the United States and many technologically advanced countries, this spans twenty-five to thirty years. The span would include the expected length of an individual professional career with an overlap on either end for upload (learning) and download (retirement) of a life's work. Individuals may have several professions in their career. Long-term solutions should address incentives to ensure their life's work is preserved. There will be multiple personal and professional data migrations in their lives. Errors will occur (lost and corrupted files) and data will not be copied absolutely correctly (formatting differences, lost elements such as footnotes, tables and images). Long-term solutions must address whether these losses are important enough to act. Finally, a trans-generational context spans multiple generations. It would include transferring data from retired researchers to an archive or repository which could withstand significant disruptions. Data should be stored such that it can be sustained beyond business cycles, the lives of people, corporations, nations and even societies and that the data could be retrieved with no knowledge of source, content, or format. This emphasizes the need for the data creator to create a description of the data, its desired use and the context in which it was collected as meta-data.

Short and long-term

- Create communications engagement strategies with individual research projects as the focus. Most research projects only last a few years. Guidelines developed for NSF DMP compliance can be used to formulate general guidelines for all research.

There is a dearth of data preservation knowledge, but a desire and motivation to learn by research faculty. Researchers do not typically seek information from the organization. However, NSF DMP guidelines, vetted by researchers and trained staff through an iterative, transparent process adds legitimacy to the information. If DMP requirements do influence behavior, campus resources can provide support for their needs. Unfortunately, as the research has shown, the professional staff is disconnected from the research community. A goal of the professional staff should be to reach out and build relationships and trust with faculty across departments that can be connected via the shared experience of a successful submission to common funding agency who have or will soon have a data management plan requirement. Identify leading researchers who can be connected through their grant activity to introduce new information. These are likely to be opinion leaders, who are crucial to disseminating the guidelines, recommending the training workshops and templates. Professional staff should continue to work with research faculty to improve implementation standards, strengthening the standards, their relationships, and trust.

Elaborating this idea in more detail, a small group could be assembled with a representative from the university repository and one college level staff person representing each department connected to NSF at the bottom of figure 16 on page 74. The college level staff person should also support research so that they have an understanding of the issues. Using the repository representative's knowledge of data preservation and the college representative's knowledge of the research and researchers, have the group discuss and create a few detailed scenarios for how data could be managed and preserved in accordance with NSF DMP requirements. Invite the three researchers represented in figure 16 to discuss their implementation plans and compare them to the scenarios proposed by the group. Propose

guidelines mutually agreed upon by the group and researchers. Send these guidelines out to everyone from the university with an NSF grant for further feedback. Finally, post the draft guidelines publically and request for comments (much like the RFC's from early Internet development) continuing refinement.

- Create training plans

The initial focus should be on meta-data creation, which was a particular weakness for everyone in the study. Professional information science and technology staff should also work directly with federal government agencies with high degree of centrality to campus, to build a framework that is useful immediately and anticipates future data access requirements. Focus staff resources on relevant agencies to improve credibility and build trust with the research community.

- Generate insights, conduct training and create templates for faculty
 - Reach out to faculty with specific workshops germane to their research
 - Schedule general training for graduate students at initial uptake, during coursework and at completion of program
- Hold mandatory workshops for all graduate students as part of initial training programs.

Although not part of the study, graduate assistants conduct the bulk of the actual research in many labs as part of their graduate programs. The training would improve the quality of lab data, improve the training, and lay the foundation for their own research when they leave the university.

- Create training opportunities for campus information technology managers and system administrators for research grant digital requirements and data mandates

Training should be extended to all campus professional IT staff. The training should be designed to familiarize them with data preservation concepts in order to build understanding and a common vocabulary. The knowledge will help IT staff provide appropriate resources to the research community. It will also help them decide when it is time to stop providing a legacy resource and provide leverage to move laggards (Rogers, 2003) to the next technology. Finally, a common framework and language can improve trust among professional IT staff and campus research groups.

Trans-generational

- Include a statement in campus strategic initiatives that explicitly sets digital integrity and data preservation goals. “Encourage and improve access to campus research through training, staff and systems to maintain research data through the next century”

There should be a statement at the institutional level that explains the strategic vision of the university with respect to digital data preservation. The statement’s scope should be broad. It should demonstrate the campus leadership acknowledgment that to exist as a 21st century research university, there must be an institutional commitment to best practices for the stewardship of digital assets. Over the past twenty-five years, the institution has built a virtual infrastructure that is as impressive as its physical infrastructure. The virtual infrastructure, by definition, is invisible and only receives attention when something bad happens. Caring for the virtual infrastructure is a complex and shared task in many ways different than physical assets. It also needs more constant care. The virtual institutional assets, digitized information generated by university personnel, are the core of the University’s mission to collect, analyze, and

disseminate knowledge. This strategic goal would make a powerful statement about the importance of institutional data stewardship.

CHAPTER 6 CONCLUSION

The project combined existing approaches and theories to study an emerging issue in a new way. The act of preserving our digital heritage is a recent phenomenon. This transformation in our era is analogous to the transition from the oral to the written traditions. Political, legal and economic systems need to adapt to new realities driven by a digital environment. The transition is occurring very quickly with little assessment of risks to long-term information needs. This supports McLuhan's claim of a societal bias towards space-bound communication. Digital data preservation problems are a result of technology innovations driven by the need to communicate faster and at greater distances. The speed of the transformation can be characterized by Moore's law, which describes the speed of innovation in the digital age. The transformation is a societal issue filled with unseen risks. The structure and components of the virtual world, by its own definition, are invisible in the physical sense. This is very different from the written tradition which has left physical artifacts for over two thousand years and that can be directly interpreted by research after years of inattention. Since we, as a society, have been very successful at finding information in our archives, we have not dealt with the problem and are passing it on to the next generation. This is a short-sighted and dangerous.

Long-term preservation of digital data is an emerging issue that demands immediate attention. Everyone bears some responsibility during the transformation. Researchers at academic institutions can be exemplars for practice, standards and behaviors. The researcher found that while there is reason for concern for the state of existing research, there is some good news. The federal government has stepped in to give central guidance by requiring researchers

to submit a data management plan. The policies that have been and will be put into place will force a cultural shift on the conduct of science and research.

It is also encouraging to note that research scientists show that they accept the new policies and some have started depositing data into repositories. It is mutually beneficial that the organization be engaged in the effort to motivate researchers towards good preservation behavior. Information scientists and technologists will need to set clear and transparent expectations for the storage with respect to short-term, long-term and trans-generational preservation.

Given the accessibility of information stored online, it is not critical that data be stored by the institution. The data will probably exist in one or more organizations. The researcher expects that identical data sets will be co-located on local, organizational, or cloud storage. The choice will be based on ease of access, storage capability, and cost. It is infinitely more important that the data be stored in such a way that it can be found in its original form fifty years from now. Thus, applying standards developed and learned through policies such as the NSF Data Management Plan requirement can and should be applied to most digital data.

Creating an adequate description of data sets with metadata was identified as an immediate critical need. However, creating metadata can be confusing and time consuming. The professional staff at the institution can assist researchers with the task. However, for this to occur, research faculty will need to deviate from normal behavior and look to the institution for guidance. The institution can facilitate this by using social network analysis to find shared activities between the campus research communities and funding sources. Professional staff can begin working with small groups of scientists from different disciplines who share funding sources. These connections can help build trust among participants and stronger reliance on

institutional resources. The researcher expects the open access that the federal government now requires will require scientists to standardize digital data preservation techniques such as consistent file naming and meta-data creation through peer pressure from their own disciplines. The shared environment fosters rapid convergence on single solutions and mutually agreed upon standards. It is important for the organization to take steps to ensure that it is explicitly engaged in the effort to motivate researcher behavior in a mutually beneficial manner.

The goal of this study was to find an approach to analyze how digital preservation could be communicated to researchers to improve the permanency of their data. The study used Diffusion of Innovations theory with Activity Theory structure and the OAIS standard to describe and analyze the problem. Using the approach, the study did improve knowledge and made recommendations for the institution.

Limitations

There are several limitations to the study. There is a general concern that the case study methodology lacks rigor, particularly when compared to other forms of inquiry (Yin, 2009, p. 14). Also, the study group is small and located at Colorado State University, so results cannot be generalized. There is also the possibility of the observer prematurely reaching conclusions. The researcher made great effort to consider the data without preconceived ideas or conclusions, particularly with the researcher's familiarity with this segment of the institution. The study method is not simple, and a little unorthodox in the way it brought together multiple data collection methods using a complex model.

Future research

The research leaves questions unanswered. Metadata is absolutely critical to preserve and retrieve data reliably. What is being done to improve metadata knowledge and practice?

Are there essential elements of information in metadata that should be included? Another question should ask what professional associations are doing to improve data preservation. Faculty are driven by the rules of their discipline, yet many in the study were not aware of any clear standards for something as simple as saving a file. The transformation from writing to digital challenges assumptions and exposes us to new vulnerabilities. Copyright and open access are the antithesis of each other and they co-exist with expected consequences today. What impact does this have on researcher willingness to share data?

Preserved data are expensive to maintain. What motivates people and organizations to maintain data that supports an unpopular view? Censorship has been largely an act of commission historically. In the new paradigm, it can be carried out as an act of omission quite easily. What does this say about censorship in the future? As cyber-warfare and cyber-crime expand and become state-sponsored, what is the threat to our stores of knowledge? Attacks carried out by a foreign government, terrorists, or disgruntled employees, as was executed against Sony in late 2014, can create substantial damage to an organization. What can be done to mitigate these incidents in a highly connected world with innumerable complex dependencies?

Finally, what is being done to promote trans-generational preservation? Just as we depend on information that is greater than a generation old, future generations will probably be as interested in how we solved our problems as the solution itself. It is certain that captains of British ships were not collecting data to support global warming computer models 250 years ago. What can we do now to assure our descendants will be able to understand our period of time? Perhaps, after a while, printed documents will become the “gold-standard” for data preservation. Perhaps we seek ways to automate selection of a class of information each generation decides

needs to be printed and stored safely. We shouldn't have to suffer a tragedy like that of the library of Alexandria to compel us to face this final problem.

BIBLIOGRAPHY

- Alfred, R. (2008, 11-04-08). Nov. 4, 1952: Univac Gets Election Right, But CBS Balks. Blog Retrieved from http://www.wired.com/science/discoveries/news/2008/11/dayintech_1104
- Bastian, M. H., S Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. Retrieved from <http://gephi.github.io/users/publications/>
- Berger, S. (2009). The Evolving Ethics of Preservation: Redefining Practices and Responsibilities in the 21st Century. *The Serials Librarian*, 57(1-2), 57-68. doi: 10.1080/03615260802669086
- Berman, F. (2008). Got data?: a guide to data preservation in the information age. *Commun. ACM*, 51(12), 50-56. doi: 10.1145/1409360.1409376
- Berners-Lee, T., & Fischetti, M. (1999). *Weaving the Web : the original design and ultimate destiny of the World Wide Web by its inventor* (1st ed.). San Francisco: HarperSanFrancisco.
- Bezos, J. (2006 August 25, 2006). Amazon Web Services Blog. from http://aws.typepad.com/aws/2006/08/amazon_ec2_beta.html
- Bricklin, D. (2012). VisiCalc: Information from its creators, Dan Bricklin and Bob Frankston. from <http://www.bricklin.com/visicalc.htm>
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 107-117. doi: [http://dx.doi.org/10.1016/S0169-7552\(98\)00110-X](http://dx.doi.org/10.1016/S0169-7552(98)00110-X)
- Carr, N. G. (2011). *The shallows : what the Internet is doing to our brains* (Norton pbk. ed. ed.). New York: W.W. Norton.
- CERN. (2012). Welcome to info.cern.ch - The website of the world's first-ever web server. Retrieved 12/29/2012, 2012, from <http://info.cern.ch/>
- Chebykin, A. I. A., Bedny, G. Z., & Karwowski, W. (2008). *Ergonomics and psychology : developments in theory and practice*. Boca Raton: CRC Press/Taylor & Francis Group.
- Chen, B. X. (2011). 4th Time a Charm for Apple? From iDisk to .Mac to MobileMe to iCloud. *Wired*. Retrieved 9/26/2015, 2015
- Child, M. S. (1992). Selection for Preservation. *Advances in Preservation and Access*(1), 212.
- Cointet, J.-P. (2012). Knowledge Community, Semantic Landscapes. 2013, from http://jph.cointet.free.fr/wp/?page_id=8

- Colorado State University Vice President for Research. (2015). Vice President for Research Data Center. Retrieved 3/20, 2015, from <http://web.research.colostate.edu/datacenter/>
- Consultative Committee Space Data Systems. (2002). *Reference model for an Open Archival Information System (OAIS) [electronic resource]*. (CCSDS 650.0-B-1 Blue Book January 2002). Washington: Washington, D.C. : CCSDS Secretariat, 2002. Retrieved from <ftp://nssdcftp.gsfc.nasa.gov/standards/nost/isoas/int07/CCSDS-650.0-W-4.pdf>.
- Conway, P. (1996). Preservation in the Digital World. Retrieved 3/29, 2009, from <http://www.clir.org/pubs/reports/conway2/>
- Council of the Consultative Committee for Space Data Systems. (2011). *Requirements For Bodies Providing Audit And Certification Of Candidate Trustworthy Digital Repositories*. (CCSDS 652.1-M-1). Washington, D.C.: CCSDS Secretariat.
- Council on Foreign Relations. (2014). Global Conflict Tracker. from CFR.org
- Cummings, L. L. B., Philip. (1996). The Organizational Trust Inventory (OTI): Development and Validation. Trust in Organizations: Frontiers of Theory and Research. SAGE Publications, Inc. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in organizations frontiers of theory and research* (pp. 302-331). Thousand Oaks, CA: SAGE Publications, Inc.
- Dimitrova, D. V., & Bugeja, M. (2007). Raising the Dead: Recovery of Decayed Online Citations. *American Communication Journal*, 9(2), 8-8.
- Ellul, J. (1964). *The technological society* ([1st American ed.]). New York,: Knopf.
- EMC2. (2014). The Digital Universe of Opportunities. *EMC Digital Universe study – with research and analysis by IDC*. 7. 2015, from <http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf>
- Engeström, Y., Miettinen, R., & Punamäki-Gitai, R.-L. (1999). *Perspectives on activity theory*. Cambridge ; New York: Cambridge University Press.
- Friese, S. (2011). *Qualitative data analysis with atlas.ti*. Thousand Oaks, CA: Sage Publications.
- Giddens, A. (1984). *The constitution of society : outline of the theory of structuration*. Cambridge Cambridgeshire: Polity Press.
- Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 1360-1380.
- Harrop, P. (2015). 3D printed car will be on the road by the end of 2015. *Electronics Weekly*(2619), 9-9.
- Hedstrom, M. (1997). Digital Preservation: A Time Bomb for Digital Libraries. *Language Resources and Evaluation*, 31(3), 189-202. doi: citeulike-article-id:2895187

- Heller-Roazen, D. (2002). Tradition's Destruction: On the Library of Alexandria. *October*, 100, 133-153. doi: 10.2307/779096
- Hjørland, B. (1997). *Information seeking and subject representation : an activity-theoretical approach to information science*. Westport, Conn.: Greenwood Press.
- IDC. (2007). The growth of the digital universe. from <http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf>
- Innis, H. A. (1972). *Empire and communications*. Toronto: University of Toronto Press.
- Institutional Research - Colorado State University. (2015). Institutional Research. Retrieved 01/14, 2014, from <http://www.ir.colostate.edu/>
- Institutional Research - Colorado State University (2012). 2012 - 2013 Fact Book. Retrieved 12/27/2012, 2012, from http://www.ir.colostate.edu/pdf/fbk/1213/2012_13_Fact_Book.pdf
- ISO. (2003). ISO 14721:2003 Space data and information transfer systems -- Open archival information system -- Reference model. Retrieved 03-18-2012, from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683
- Jesdanun, A. (2003). Digital Memory Threatened as File Formats Evolve. Retrieved 02/01/12, from <http://www.crn.com/news/storage/18838546/digital-memory-threatened-as-file-formats-evolve.htm>
- Kramer, R. M., & Tyler, T. R. (1996). *Trust in organizations : frontiers of theory and research*. Thousand Oaks, Calif.: Sage Publications.
- Kuny, T. (1997). *A Digital Dark Ages? Challenges in the Preservation of Electronic Information*. Paper presented at the 63rd IFLA Council and General Conference, Copenhagen, Denmark. <http://www.ifla.org/IV/ifla63/63kuny1.pdf>
- Lee, C. A. (2005). *Defining digital preservation work: a case study of the development of the reference model for an open archival information system*.
- Library of Congress. (2007a). Sustainability Factors. Retrieved 03-12-10, 2011, from <http://www.digitalpreservation.gov/formats/sustain/sustain.shtml>
- Library of Congress. (2007b, 03/ 7/2007). Sustainability of Digital Formats Planning for Library of Congress Collections. from http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml
- Licklider, J. C. R. (1960). Man-computer symbiosis. *Human Factors in Electronics, IRE Transactions on*(1), 4-11.

- Logan, R. K. (2004). *The alphabet effect : a media ecology understanding of the making of Western civilization*. Cresskill, N.J.: Hampton Press.
- McLuhan, M. (1964). *Understanding media; the extensions of man* ([1st ed.]). New York,: McGraw-Hill.
- Milgram, S. (1967). The small world problem. *Psychology today*, 2(1), 60-67.
- Montagne, R. (Writer). (2008). NPR's business news starts with video games going political [Radio], *Morning Edition*. US: NPR.
- Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8), 4.
- Moore, G. E. (1975). Progress in digital integrated electronics. *IEDM Tech. Digest*, 11.
- Morgan Library - Colorado State University. (2012). About the CSU Digital Repository. Retrieved 12/27/2012, 2012, from <http://lib.colostate.edu/repository/>
- Morten Fjeld , K. L., Martin Bichsel, Fred Voorhorst, Helmut Krueger and Matthias Rauterberg. (2002). Physical and Virtual Tools: Activity Theory Applied to the Design of Groupware *Computer Supported Cooperative Work (CSCW)*, 11(1-2), 28. doi: 10.1023/A:1015269228596
- Mumford, L. (1974). *The pentagon of power*. New York,: Harcourt Brace Jovanovich.
- Nardi, B. A. (1996). *Context and consciousness : activity theory and human-computer interaction*. Cambridge, Mass.: MIT Press.
- National Institutes of Health. (2003). Final NIH Statement On Sharing Research Data. 2012, from <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>
- National Science Foundation. (2011a). Data Management for NSF BIO Directorate Proposals and Awards. from <http://www.nsf.gov/bio/pubs/BIODMP061511.pdf>
- National Science Foundation. (2011b). Data Management for NSF EHR Directorate Proposals and Awards. from <http://www.nsf.gov/bfa/dias/policy/dmpdocs/ehr.pdf>
- National Science Foundation. (2011c). Data Management for NSF SBE Directorate Proposals and Awards. from http://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf
- National Science Foundation. (2012a). Data Management & Sharing Frequently Asked Questions (FAQs). Retrieved 03/27/2012, 2012, from <http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp#1>
- National Science Foundation. (2012b). Requirements by Directorate, Office, Division, Program, or other NSF Unit. *Dissemination and Sharing of Research Results*. Retrieved 03/27/2012, 2012, from <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>

- National Science Foundation. (2013). Grant Proposal Guide - Chapter II - Proposal Preparation Instructions. Retrieved 04/13, 2015, from http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp
- New York Stock Exchange. (2015). NYSEData.com Factbook: NYSE overview statistics. from http://www.nyxdata.com/nysedata/asp/factbook/viewer_edition.asp?mode=table&key=268&category=14
- Office of Management and Budget. (1999). *Uniform administrative requirements for grants and agreements with institutions of higher education, hospitals, and other non-profit organizations*. [Washington, D.C.]: Executive Office of the President, Office of Management and Budget.
- Office of Science and Technology Policy. (2013). *Increasing Access to the Results of Federally Funded Scientific Research*. Washington, D.C.: Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
- Office of the President. (2013). *Executive Order -- Making Open and Machine Readable the New Default for Government Information*. White House: Retrieved from <https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government->.
- PARC, X. (2012). Milestones. Retrieved 03-21, 2012, from <http://www.parc.com/about/>
- Plato, & Jowett, B. (1931). *The dialogues of Plato* (3d ed.). London,: H. Milford, Oxford university press.
- Rao, L. (2012). Box: The Path From Arrington's Backyard To A Billion Dollar Business. *Techcrunch*.
- Raymond, A. L. (2001). *Long term preservation of digital information*. Paper presented at the Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, Roanoke, Virginia, United States.
- Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.
- Sanders, J. (2014). Microsoft renames SkyDrive to more confusing OneDrive amid legal complaint. *TechRepublic*. <http://www.techrepublic.com/article/microsoft-renames-skydrive-to-more-confusing-onedrive-amid-legal-complaint/>
- Sannino, A., Daniels, H., & Gutierrez, K. D. (2009). Activity theory between historical engagement and future making practice. In A. Sannino, H. Daniels & K. D. Gutierrez (Eds.), *Learning and expanding with activity theory* (pp. xxi, 367 p.). New York: Cambridge University Press
- Sapir, E. (1921). *Language, an introduction to the study of speech*. New York,: Harcourt, Brace and company.

- Sapir, E., & Mandelbaum, D. G. (1949). *Selected writings in language, culture and personality*. Berkeley,: University of California Press.
- Scott, J. (2000). *Social network analysis : a handbook* (2nd ed.). London ; Thousands Oaks, Calif.: SAGE Publications.
- Seel, P. B. (2012). *Digital universe : the global telecommunication revolution*. Malden, MA: Wiley-Blackwell.
- Spencer, L. J. a. A. A. (2012). *The Problem of Data: Data Management and Curation Practices Among University Researchers CLIR pub 154* (pp. 43). Retrieved from <http://www.clir.org/pubs/reports/pub154>
- Stern, N. B. (1981). *From ENIAC to UNIVAC : an appraisal of the Eckert-Mauchly computers*. Bedford, Mass.: Digital Press.
- Stille, A. (2006). Are We Losing Our Memory? or The Museum of Obsolete Technology. Retrieved 4/2, 2009, from <http://www.lostmag.com/issue3/memory.php>
- Suchman, L. A. (1987). *Plans and situated actions : the problem of human-machine communication*. Cambridge Cambridgeshire ; New York: Cambridge University Press.
- Sunderland, U. o. (2009 Tuesday 6th October 2009). 18th century ships' logs predict future weather forecast. Retrieved 02/01, 2012, from <http://www.sunderland.ac.uk/newsevents/news/news/index.php?nid=734>
- Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1), 230-265. doi: 10.1112/plms/s2-42.1.230
- Union, W. (1965). Statement of Company goals - March 31, 1965 (pp. 16). Smithsonian National Museum of American History, Lemuelson Center for the Study of Invention & Innovation, in the Western Union Telegraph Company Records archival collection covering the years 1820-1995. Archives Center, National Museum of American History, Smithsonian Institution, Washington, D.C.; SERIES 3: EXECUTIVE RECORDS, 1848-1987 box 196, folder 6,
- Walters, T. O. M., Robert H. (2008). *Creating Trust Relationships for Distributed Digital Preservation Federations*. Paper presented at the 5th International Conference on Preservation of Digital Objects (iPRES). London.
- Watts, D. J. (2003). *Six degrees : the science of a connected age* (1st ed.). New York: Norton.
- Weiser, M. (1991). The Computer for the Twenty-First Century. *Scientific American*, 265(3), 94-104. doi: citeulike-article-id:771482
- Wilson, T. D. (2008). Activity theory and information seeking. *Annual Review of Information Science and Technology*, 42(1), 119-161. doi: 10.1002/aris.2008.1440420111

Yin, R. K. (2009). *Case study research : design and methods* (4th ed.). Los Angeles, Calif.: Sage Publications.

APPENDICES

APPENDIX A - Initial Interview questions (Spencer, p. 26)

Introduction

1. In a research project that you are currently working on (or recently completed) narrate the data management process.
2. Did this project have a data preservation or management plan requirement?

Training

3. Have you had training in data curation?
4. Who provided the training?
5. If so, what kind/what tools?
6. Do you feel that it was adequate?
7. What would you like to know more about?

Project

8. What are the sources of data?
9. Did you create new data sources as part of this research (e.g., experimental results, data sets, coding files, indexes)?
10. How did you organize the data?
11. Did you document this system?
12. Are the data backed up?
13. How/Where?
14. Do you keep a personal archive of materials related to your scholarship (e.g., field notes, lab books, e-mails, photographs)?
15. How/Where are they stored?
16. If you wanted to go back and work with the data again, what would be the most important information to have?
17. If someone wanted to replicate/reconstruct your analysis, what information would be needed?

Collaboration

18. Do you collaborate with other researchers on this project?
19. How did you manage version control?
20. What software (if any) did you use?

Extra-departmental activities

21. Outside of your department, do you participate in any campus activity, either sanctioned or non-sanctioned, such as a campus committee, weekly cup of coffee or recreational activity with faculty not from your department?
22. Have you ever discussed preserving files during these activities? Explain.
23. Do you belong to organizations off campus in which there are other university faculty?
24. Have you ever discussed preserving digital data with anyone from off campus organizations? Explain.

Preservation

25. Once you were finished with this project, do you have a plan/strategy for archiving these materials? (What will not be held?)
26. Where will they be held?
27. Who is responsible for them?
28. (If not) Why don't you archive your materials?
29. Did anyone offer guidance in making these decisions?
30. If someone were to return to your data in 5 to 10 years (or longer), what contextual information would be needed?
31. If you were archiving your research for future scholars, what would be the most important things to be preserved?
32. Who would potentially re-use this data?
33. What are your expectations for this re-use (e.g., citation, copies of papers, reciprocity)?

APPENDIX B - Face to Face Survey Questions

Data Preservation Communication

Intro These definitions of the stages of data in the research process below are important for questions in this survey. A copy will be handed to each participant to read prior to starting the survey.

Raw - The data are newly created, generated or acquired.

Processed - The raw data is reviewed, refined or revised to better enable its use in the research. This may include reducing “noise” in the data, removing elements in the data that are superfluous to its use, or checking for errors. Processing data may also include adding additional or supplementary information including metadata to the data set.

Analyzed - The stage in which data are critically examined by the researcher(s) to provide information or answers to their research questions. The process of analyzing the data may produce new data sets, by-products, or other outputs that should be accounted for.

Finalized - The last stage in the data lifecycle in which all re-workings and manipulations of the data by the researcher have ceased.

Backup - Backup data are created during a project and is intended for disaster recovery, particularly during a project

Archive - Archived data are stored data intended for later use. Archived data are considered long-term storage

Data Repository - a central place where data are stored and maintained by professional staff.

Metadata - Information about the data such as creation date, basic description, author, etc.

1 Respondent Tracking Number

2 How would you describe your knowledge of the following?

_____ Good file naming conventions (1)

_____ Meta-data (2)

_____ Back-up strategies (3)

_____ Long-term data preservation (4)

3 What percentage of the storage media for your research data are ...?

_____ Managed by you personally (1)

_____ Managed by someone within a research / collaboration group (2)

_____ Managed by Dept./College IT staff (3)

_____ Managed by the University IT staff (4)

_____ Third Party (5)

4 How often do you use the following as part of your standard file naming convention for your research data?

- _____ Represents file contents (6)
- _____ Naturally ordered numeric or alphabetic (7)
- _____ Facilitates version control (10)
- _____ Follows standards established by your discipline (9)
- _____ Consistent throughout your research project file system (8)

5 How frequently do you use organize and describe your data so that another person with similar expertise can understand and properly use it?

_____ Your Choice (3)

6 How frequently do your research tools – software or hardware – automatically generate useful meta-data?

_____ -- (1)

7 How much of your research data are:

	0 - 25% (1)	26 - 50% (2)	51 - 75% (3)	75 - 100% (4)
Raw (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Processed (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Analyzed (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Finalized (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

8 How frequently do you review all files, including old projects, to ensure they can be found, opened and read?

	Never (1)	When I need to review old research (2)	Every Few Years (3)	Annually (4)
Raw (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Processed (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Analyzed (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Finalized (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

9 In general, how long do you believe your data sets will be useful or have value for you or others if they were to be preserved?

	My data-set does not need to be preserved (1)	0 - 5 years (2)	5 - 15 years (3)	15 - 30 years (4)	30 - 100 years (5)	Indefinitely (6)	N/A (7)
Raw (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Processed (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Analyzed (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Finalized (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

10 Rank responsibility for archiving your research data once you finish a project, (archived data is created during a project that is stored after it has been completed)?

- You personally? (1)
 Managed within a research / collaboration group (2)
 Dept./College IT staff (3)
 University IT staff (4)
 Third party (5)

11 What is the most time a data incident in which critical research data was permanently lost cost you?

- Yes (1)

12 How frequently do you make back-up copies of your data (Backup data is created during a project and is intended for disaster recovery)?

- (1)

13 The backups are typically:

- Manual (1)
 Automatic (2)

14 The backup files are:

- Managed by you personally (1)
 Managed by someone in your research / collaboration group (2)
 Managed by Dept./College IT staff (3)
 Managed by University IT staff (4)
 Managed by a third party (5)

Instr As we talk, I am going to discuss your most sensitive and least sensitive data. Think of a recent research project to answer the following four questions. For the following two questions, please consider your most sensitive research data

15 In the table below, please indicate what data you would be willing to share during your research. (Please select as many as apply)

- Raw (1)
- Processed (2)
- Analyzed (3)
- Finalized (4)

16 In the table below, please indicate your willingness to share data after project is completed and published. (Please select as many as apply)

- Raw (1)
- Processed (2)
- Analyzed (3)
- Finalized (4)

Instr For the following two questions, please consider your most sensitive research data

17 In the table below, please indicate what data you would be willing to share during your research. (Please select as many as apply)

- Raw (1)
- Processed (2)
- Analyzed (3)
- Finalized (4)

18 In the table below, please indicate your willingness to share data after project is completed and published. (Please select as many as apply)

- Raw (1)
- Processed (2)
- Analyzed (3)
- Finalized (4)

19 Select the number from the scale, 1 is lowest and 7 highest, that is closest to your opinion.

_____ My level of confidence that the University IT staff is technically competent at the critical elements of their job is (1)

_____ My level of confidence that the University IT staff will make well thought out decisions about their job is (2)

_____ My level of confidence that the University IT staff will follow through on assignments is (3)

_____ My level of confidence that the University IT staff has an acceptable level of understanding of their job is (4)

_____ My level of confidence that the University IT staff will be able to do their job in an acceptable manner is (5)

_____ When the University IT staff tells me something, my level of confidence that I can rely on what they tell me is (6)

_____ My confidence in the University IT staff to do the job without causing other problems is (7)

_____ My level of confidence that the University IT staff will think through what they are doing on the job is (8)

20 Do any of your funding sources require that you create a data management plan as a condition of funding?

Yes (1)

No (2)

21 Do any of your funding sources require that you share your data with others, publish your data, or deposit your data into a data repository?

Yes (1)

No (2)

22 Do any of your funding sources require that you preserve your data beyond the life of the funding?

Yes (1)

No (2)

23 Do any of your funding agencies require that you place your research data into a data repository (a central place where data is stored and maintained)?

Yes (1)

No (2)

24 Have you ever deposited data into a data repository (a data repository refers to a central place where data is stored and maintained)?

CSU (1)

Other (2)

25 Repository name or location:

26 If not required, would you be willing to submit your data to a data repository in the future?

- Yes (1)
- No (2)

27 Have you ever deposited any metadata into a data repository?

- CSU (Digitool) (1)
- Other (2)

28 If "other", where have you deposited meta-data into a centralized repository (may be both)?

29 If not required, would you be willing to submit your meta-data into a centralized repository in the future?

- Yes (1)
- No (2)

30 Have you ever received digital data management training?

- Yes (1)
- No (2)

31 How many data management workshops or sessions have you attended to the best of your recollection?

	In the past three years (2)	In my career (3)
CSU Hosted (1)		
Non-CSU Hosted (2)		

32 If non-CSU hosted, who hosted?

33 Please select what areas your data management training covered

- File Naming (7)
- Meta-data Creation (8)
- Back-up Strategies (9)
- Data Preservation (4)

34 Select the number from the scale, 1 is lowest and 7 highest, that is closest to your opinion.

_____ My level of confidence that the College or Departmental IT staff is technically competent at the critical elements of their job is (1)

_____ My level of confidence that the College or Departmental IT staff will make well thought out decisions about their job is (2)

_____ My level of confidence that the College or Departmental IT staff will follow through on assignments is (3)

_____ My level of confidence that the College or Departmental IT staff has an acceptable level of understanding of their job is (4)

_____ My level of confidence that the College or Departmental IT staff will be able to do their job in an acceptable manner is (5)

_____ When the College or Departmental IT staff tells me something, my level of confidence that I can rely on what they tell me is (6)

_____ My confidence in the College or Departmental IT staff to do the job without causing other problems is (7)

_____ My level of confidence that the College or Departmental IT staff will think through what they are doing on the job is (8)

APPENDIX C - Final Interview Questions

1. How should the university dedicate resource to help researchers preserve their digital research data?
 - a. Categories
 - i. Technical Staff
 - ii. Systems
 - iii. Training
2. Has/should the university considered a risk management plan/portfolio for its digital data in general?
3. If it has or should, how?

APPENDIX D - National Science Foundation Organization Chart

<u>Office of the Director</u>	OD
<u>Office of the General Counsel</u>	OD/OGC
<u>Office of International and Integrative Activities</u>	OD/IIA
<u>Office of Legislative & Public Affairs</u>	OD/OLPA
<u>Office of Diversity and Inclusion</u>	OD/ODI
<u>National Science Board</u>	NSB
<u>Office of the Inspector General</u>	OIG
<u>Directorate for Biological Sciences</u>	BIO/OAD
<u>Division of Molecular & Cellular Biosciences</u>	BIO/MCB
<u>Division of Biological Infrastructure</u>	BIO/DBI
<u>Division of Integrative Organismal Systems</u>	BIO/IOS
<u>Division of Environmental Biology</u>	BIO/DEB
<u>Emerging Frontiers Office</u>	BIO/EF
<u>Directorate for Computer & Information Science & Engineering</u>	CISE/OAD
<u>Division of Advanced Cyberinfrastructure</u>	CISE/ACI
<u>Division of Computing and Communication Foundations</u>	CISE/CCF
<u>Division of Computer and Network Systems</u>	CISE/CNS
<u>Division of Information and Intelligent Systems</u>	CISE/IIS
<u>Directorate for Education & Human Resources</u>	EHR/OAD
<u>Division of Research on Learning in Formal and Informal Settings</u>	EHR/DRL
<u>Division of Graduate Education</u>	EHR/DGE
<u>Division of Human Resource Development</u>	EHR/HRD
<u>Division of Undergraduate Education</u>	EHR/DUE

<u>Directorate for Engineering</u>	ENG/OAD
<u>Division of Chemical, Bioengineering, Environmental, and Transport Systems</u>	ENG/CBET
<u>Division of Civil, Mechanical & Manufacturing Innovation</u>	ENG/CMMI
<u>Division of Electrical, Communications & Cyber Systems</u>	ENG/ECCS
<u>Division of Engineering Education & Centers</u>	ENG/EEC
<u>Division of Industrial Innovation & Partnerships</u>	ENG/IIP
<u>Office of Emerging Frontiers in Research & Innovation</u>	ENG/EFRI
<u>Directorate for Geosciences</u>	GEO/OAD
<u>Division of Atmospheric and Geospace Sciences</u>	GEO/AGS
<u>Division of Earth Sciences</u>	GEO/EAR
<u>Division of Ocean Sciences</u>	GEO/OCE
<u>Division of Polar Programs</u>	GEO/PLR
<u>Directorate for Mathematical & Physical Sciences</u>	MPS/OAD
<u>Division of Astronomical Sciences</u>	MPS/AST
<u>Division of Chemistry</u>	MPS/CHE
<u>Division of Materials Research</u>	MPS/DMR
<u>Division of Mathematical Sciences</u>	MPS/DMS
<u>Division of Physics</u>	MPS/PHY
<u>Directorate for Social, Behavioral & Economic Sciences</u>	SBE/OAD
<u>Division of Social and Economic Sciences</u>	SBE/SES
<u>Division of Behavioral and Cognitive Sciences</u>	SBE/BCS
<u>National Center for Science and Engineering Statistics</u>	SBE/NCSE
<u>SBE Office of Multidisciplinary Activities</u>	SBE/SMA
<u>Office of Budget, Finance, and Award Management</u>	BFA/OAD
<u>Budget Division</u>	BFA/BD
<u>Division of Acquisition and Cooperative Support</u>	BFA/DACS
<u>Division of Grants & Agreements</u>	BFA/DGA
<u>Division of Financial Management</u>	BFA/DFM
<u>Division of Institution and Award Support</u>	BFA/DIAS
<u>Large Facilities Office</u>	BFA/LFO
<u>Office of Information & Resource Management</u>	OIRM/OAD
<u>Division of Human Resources Management</u>	OIRM/HRM
<u>Division of Information Systems</u>	OIRM/DIS
<u>Division of Administrative Services</u>	OIRM/DAS



Grant Proposal Guide

NSF 11-1 January 2011

CHAPTER II - PROPOSAL PREPARATION INSTRUCTIONS

Each proposing organization that is new to NSF or has not received an NSF grant within the previous two years should be prepared to submit basic organization and management information and certifications, when requested, to the applicable award making division within BFA. The requisite information is described in the Prospective New Awardee Guide. The information contained in this Guide will assist the organization in preparing documents which the National Science Foundation requires to conduct administrative and financial reviews of the organization. This Guide also serves as a means of highlighting the accountability requirements associated with Federal awards.

To facilitate proposal preparation, Frequently Asked Questions (FAQs) regarding proposal preparation and submission are available electronically on the NSF website.⁹

A. Conformance with Instructions for Proposal Preparation

It is important that all proposals conform to the instructions provided in the GPG. Conformance is required and will be strictly enforced unless an authorization to deviate from standard proposal preparation requirements has been approved. NSF may return without review proposals that are not consistent with these instructions. See GPG Chapter IV.B for additional information. NSF must authorize any deviations from these instructions in advance of proposal submission. Deviations may be authorized in one of two ways:

1. through specification of different requirements in an NSF program solicitation; or
2. by the written approval of the cognizant NSF Assistant Director/Office Head or designee. These approvals to deviate from NSF proposal preparation instructions may cover a particular program or programs or, in rare instances, an “individual” deviation for a particular proposal.

Proposers may deviate from these instructions only to the extent authorized. Proposals must include an authorization to deviate from standard NSF proposal preparation instructions has been received in one of the following ways, as appropriate: (a) by identifying the solicitation number that authorized the deviation in the appropriate block on the proposal Cover Sheet; or (b) for individual deviations, by identifying the name, date and title of the NSF official authorizing the deviation.¹⁰ Further instructions are available on the FastLane website.

B. Format of the Proposal

Prior to electronic submission, it is strongly recommended that proposers conduct an administrative review to ensure that proposals comply with the proposal preparation guidelines established in the GPG. GPG Exhibit II-1 contains a proposal preparation checklist that may be used to assist in this review. This checklist is not intended to be an all-inclusive repetition of the required proposal contents and associated proposal preparation guidelines. It is, however, meant to highlight certain critical items so they will not be overlooked when the proposal is prepared.

1. Proposal Pagination Instructions considered part of the 15-page Project Description limitation. This Special Information and Supplementary Documentation section also is not considered an appendix. Specific guidance on the need for additional documentation may be obtained from the organization's sponsored projects office or in the references cited below.

- **Postdoctoral Researcher Mentoring Plan.** Each proposal³³ that requests funding to support postdoctoral researchers³⁴ must include, as a supplementary document, a description of the mentoring activities that will be provided for such individuals. In no more than one page, the mentoring plan must describe the mentoring that will be provided to all postdoctoral researchers supported by the project, irrespective of whether they reside at the submitting organization, any subawardee organization, or at any organization participating in a simultaneously submitted collaborative project. Proposers are advised that the mentoring plan may not be used to circumvent the 15-page project description limitation. See GPG Chapter II.D.4 for additional information on collaborative proposals.

Examples of mentoring activities include, but are not limited to: career counseling; training in preparation of grant proposals, publications and presentations; guidance on ways to improve teaching and mentoring skills; guidance on how to effectively collaborate with researchers from diverse backgrounds and disciplinary areas; and training in responsible professional practices. The proposed mentoring activities will be evaluated as part of the merit review process under the Foundation's broader impacts merit review criterion.

- **Plans for data management and sharing of the products of research.** Proposals must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (see AAG Chapter VI.D.4), and may include:
 1. the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
 2. the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);

3. policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
4. policies and provisions for re-use, re-distribution, and the production of derivatives;
and
5. plans for archiving data, samples, and other research products, and for preservation of access to them.

Data management requirements and plans specific to the Directorate, Office, Division, Program, or other NSF unit, relevant to a proposal are available at: <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>. If guidance specific to the program is not available, then the requirements established in this section apply.

Simultaneously submitted collaborative proposals and proposals that include subawards are a single unified project and should include only one supplemental combined Data Management Plan, regardless of the number of non-lead collaborative proposals or subawards included. Fastlane will not permit submission of a proposal that is missing a Data Management Plan. Proposals for supplementary support to an existing award are not required to include a Data Management Plan.

A valid Data Management Plan may include only the statement that no detailed plan is needed, as long as the statement is accompanied by a clear justification. Proposers who feel that the plan cannot fit within the supplement limit of two pages may use part of the 15- page Project Description for additional data management information. Proposers are advised that the Data Management Plan may not be used to circumvent the 15-page Project Description limitation. The Data Management Plan will be reviewed as an integral part of the proposal, coming under Intellectual Merit or Broader Impacts or both, as appropriate for the scientific community of relevance.

APPENDIX F - Open Data Policy - Office of the President of the United States



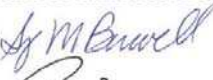
THE DIRECTOR

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

May 9, 2013

M-13-13

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Sylvia M. Burwell 
Director

Steven VanRoekel 
Federal Chief Information Officer

Todd Park 
U.S. Chief Technology Officer

Dominic J. Mancini 
Acting Administrator, Office of Information and Regulatory Affairs

SUBJECT: Open Data Policy—Managing Information as an Asset

Information is a valuable national resource and a strategic asset to the Federal Government, its partners, and the public. In order to ensure that the Federal Government is taking full advantage of its information resources, executive departments and agencies (hereafter referred to as “agencies”) must manage information as an asset throughout its life cycle to promote openness and interoperability, and properly safeguard systems and information. Managing government information as an asset will increase operational efficiencies, reduce costs, improve services, support mission needs, safeguard personal information, and increase public access to valuable government information.

Making information resources accessible, discoverable, and usable by the public can help fuel entrepreneurship, innovation, and scientific discovery – all of which improve Americans’ lives and contribute significantly to job creation. For example, decades ago, the Federal Government made both weather data and the Global Positioning System (GPS) freely available to anyone. Since then, American entrepreneurs and innovators have used these resources to create navigation systems, weather newscasts and warning systems, location-based applications, precision farming tools, and much more.

Pursuant to Executive Order of May 9, 2013, *Making Open and Machine Readable the New Default for Government Information*, this Memorandum establishes a framework to help institutionalize the principles of effective information management at each stage of the information’s life cycle to promote interoperability and openness. Whether or not particular information can be made public, agencies can apply this framework to all information resources to promote efficiency and produce value.

Specifically, this Memorandum requires agencies to collect or create information in a way that supports downstream information processing and dissemination activities. This includes using machine-readable and open formats, data standards, and common core and extensible metadata for all new

information creation and collection efforts. It also includes agencies ensuring information stewardship through the use of open licenses and review of information for privacy, confidentiality, security, or other restrictions to release. Additionally, it involves agencies building or modernizing information systems in a way that maximizes interoperability and information accessibility, maintains internal and external data asset inventories, enhances information safeguards, and clarifies information management responsibilities.

The Federal Government has already made significant progress in improving its management of information resources to increase interoperability and openness. The President's Memorandum on *Transparency and Open Government*¹ instructed agencies to take specific actions to implement the principles of transparency, participation, and collaboration, and the Office of Management and Budget's (OMB) *Open Government Directive*² required agencies to expand access to information by making it available online in open formats. OMB has also developed policies to help agencies incorporate sound information practices, including OMB Circular A-130³ and OMB Memorandum M-06-02.⁴ In addition, the Federal Government launched Data.gov, an online platform designed to increase access to Federal datasets. The publication of thousands of data assets through Data.gov has enabled the development of numerous products and services that benefit the public.

To help build on these efforts, the President issued a Memorandum on May 23, 2012 entitled *Building a 21st Century Digital Government*⁵ that charged the Federal Chief Information Officer (CIO) with developing and implementing a comprehensive government-wide strategy to deliver better digital services to the American people. The resulting *Digital Government Strategy*⁶ outlined an information-centric approach to transform how the Federal Government builds and delivers digital services, and required OMB to develop guidance to increase the interoperability and openness of government information.

This Memorandum is designed to be consistent with existing requirements in the Paperwork Reduction Act,⁷ the E-Government Act of 2002,⁸ the Privacy Act of 1974,⁹ the Federal Information Security Management Act of 2002 (FISMA),¹⁰ the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA),¹¹ the Freedom of Information Act,¹² the Information Quality Act,¹³ the

¹ President Barack Obama, Memorandum on Transparency and Open Government (Jan. 21, 2009), available at http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment.

² OMB Memorandum M-10-06, *Open Government Directive* (Dec. 8, 2009), available at http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf

³ OMB Circular A-130, available at http://www.whitehouse.gov/omb/Circulars_a130_a130trans4/

⁴ OMB Memorandum M-06-02, *Improving Public Access to and Dissemination of Government Information and Using the Federal Enterprise Architecture Data Reference Model* (Dec. 16, 2005), available at <http://www.whitehouse.gov/sites/default/files/omb/memoranda/fy2006/m06-02.pdf>

⁵ President Barack Obama, Memorandum on Building a 21st Century Digital Government (May 23, 2012), available at http://www.whitehouse.gov/sites/default/files/uploads/2012digital_mcm_rel.pdf

⁶ Office of Management and Budget, *Digital Government: Building a 21st Century Platform to Better Serve the American People* (May 23, 2012), available at <http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government-strategy.pdf>

⁷ 44 U.S.C. § 3501 *et seq.*

⁸ Pub. L. No. 107-347, 116 Stat. 2899 (2002) (codified as 44 U.S.C. § 3501 note).

⁹ 5 U.S.C. § 552a.

¹⁰ 44 U.S.C. § 3541, *et seq.*

¹¹ Section 503(a), Pub. L. No. 107-347, 116 Stat. 2899 (2002) (codified as 44 U.S.C. § 3501 note); *see also* Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), available at http://www.whitehouse.gov/sites/default/files/omb/assets/omb/fedreg/2007/061507_cipsea_guidance.pdf

¹² 5 USC 552(a)(2).

Federal Records Act,¹⁴ and existing OMB and Office of Science and Technology Policy (OSTP) guidance.

If agencies have any questions regarding this Memorandum, please direct them to OMB at datase@omb.eop.gov.

Attachment

¹³ Pub. L. No. 106-554 (2000) (codified at 44 U.S.C. § 3504(d)(1) and 3516). See also OMB Memorandum M-12-18, *Managing Government Records Directive* (Aug. 24, 2012), available at <http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-18.pdf>.

¹⁴ 44 U.S.C. Chapters 21, 22, 29, 31, and 33. See also 36 CFR Subchapter B - Records Management.

Attachment

This attachment provides definitions and implementation guidance for M-13-13, *Open Data Policy—Managing Information as an Asset*.

I. Definitions:

Data: For the purposes of this Memorandum, the term “data” refers to all structured information, unless otherwise noted.¹⁵

Dataset: For the purposes of this Memorandum, the term “dataset” refers to a collection of data presented in tabular or non-tabular form.

Fair Information Practice Principles: The term “Fair Information Practice Principles” refers to the eight widely accepted principles for identifying and mitigating privacy impacts in information systems, programs and processes, delineated in the National Strategy for Trusted Identities in Cyberspace.¹⁶

Government information: As defined in OMB Circular A-130, “government information” means information created, collected, processed, disseminated, or disposed of, by or for the Federal Government.

Information: As defined in OMB Circular A-130, the term “information” means any communication or representation of knowledge such as facts, data, or opinions in any medium or form, including textual, numerical, graphic, cartographic, narrative, or audiovisual forms.

Information life cycle: As defined in OMB Circular A-130, the term “information life cycle” means the stages through which information passes, typically characterized as creation or collection, processing, dissemination, use, storage, and disposition.

Personally identifiable information: As defined in OMB Memorandum M-10-23,¹⁷ “personally identifiable information” (PII) refers to information that can be used to distinguish or trace an individual’s identity, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual. The definition of PII is not anchored to any single category of information or technology. Rather, it requires a case-by-case assessment of the specific risk that an individual can be identified. In performing this assessment, it is important for an agency to recognize that non-PII can become PII whenever additional information is made publicly available (in any medium and from any source) that, when combined with other available information, could be used to identify an individual.

Mosaic effect: The mosaic effect occurs when the information in an individual dataset, in isolation, may not pose a risk of identifying an individual (or threatening some other important interest such as security), but when combined with other available information, could pose such risk. Before disclosing potential PII or other potentially sensitive information, agencies must consider other publicly available data – in

¹⁵ *Structured* information is to be contrasted with *unstructured* information (commonly referred to as “content”) such as press releases and fact sheets. As described in the *Digital Government Strategy*, content may be converted to a structured format and treated as data. For example, a web-based fact sheet may be broken into the following component data pieces: the title, body text, images, and related links.

¹⁶ The White House, *National Strategy for Trusted Identities in Cyberspace* (April 2011), available at http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf

¹⁷ OMB Memorandum M-10-23, *Guidance for Agency Use of Third-Party Websites and Applications* (June 25, 2010), available at http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-23.pdf

any medium and from any source – to determine whether some combination of existing data and the data intended to be publicly released could allow for the identification of an individual or pose another security concern.

Open data: For the purposes of this Memorandum, the term “open data” refers to publicly available data structured in a way that enables the data to be fully discoverable and usable by end users. In general, open data will be consistent with the following principles:

- *Public.* Consistent with OMB’s *Open Government Directive*, agencies must adopt a presumption in favor of openness to the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions.
- *Accessible.* Open data are made available in convenient, modifiable, and open formats that can be retrieved, downloaded, indexed, and searched. Formats should be machine-readable (i.e., data are reasonably structured to allow automated processing). Open data structures do not discriminate against any person or group of persons and should be made available to the widest range of users for the widest range of purposes, often by providing the data in multiple formats for consumption. To the extent permitted by law, these formats should be non-proprietary, publicly available, and no restrictions should be placed upon their use.
- *Described.* Open data are described fully so that consumers of the data have sufficient information to understand their strengths, weaknesses, analytical limitations, security requirements, as well as how to process them. This involves the use of robust, granular metadata (i.e., fields or elements that describe data), thorough documentation of data elements, data dictionaries, and, if applicable, additional descriptions of the purpose of the collection, the population of interest, the characteristics of the sample, and the method of data collection.
- *Reusable.* Open data are made available under an open license that places no restrictions on their use.
- *Complete.* Open data are published in primary forms (i.e., as collected at the source), with the finest possible level of granularity that is practicable and permitted by law and other requirements. Derived or aggregate open data should also be published but must reference the primary data.
- *Timely.* Open data are made available as quickly as necessary to preserve the value of the data. Frequency of release should account for key audiences and downstream needs.
- *Managed Post-Release.* A point of contact must be designated to assist with data use and to respond to complaints about adherence to these open data requirements.

Project Open Data: “Project Open Data,” a new OMB and OSTP resource, is an online repository of tools, best practices, and schema to help agencies adopt the framework presented in this guidance. Project Open Data can be accessed at <http://project-open-data.github.io>.¹⁸ Project Open Data will evolve over time as a community resource to facilitate adoption of open data practices. The repository includes definitions, code, checklists, case studies, and more, and enables collaboration across the Federal Government, in partnership with public developers, as applicable. Agencies can visit Project Open Data for a more comprehensive glossary of terms related to open data.

¹⁸ Links to the best practices developed in Project Open Data referenced in this memorandum can be found through the directory on this main page.

II. Scope:

The requirements in part III, sections 1 and 2 of this Memorandum apply to all new information collection, creation, and system development efforts as well as major modernization projects that update or re-design existing information systems. National Security Systems, as defined in 40 U.S. C. 11103, are exempt from the requirements of this policy. The requirements in part III, section 3 apply to management of all datasets used in an agency's information systems. Agencies are also encouraged to improve the discoverability and usability of existing datasets by making them "open" using the methods outlined in this Memorandum, prioritizing those that have already been released to the public or otherwise deemed high-value or high-demand through engagement with customers (see part III, section 3.c). Agencies should exercise judgment before publicly distributing data residing in an existing system by weighing the value of openness against the cost of making those data public.

III. Policy Requirements:

Agencies management of information resources must begin at the earliest stages of the planning process, well before information is collected or created. Early strategic planning will allow the Federal Government to design systems and develop processes that unlock the full value of the information, and provide a foundation from which agencies can continue to manage information throughout its life cycle.

Agencies shall take the following actions to improve the management of information resources throughout the information's life cycle and reinforce the government's presumption in favor of openness:

1. **Collect or create information in a way that supports downstream information processing and dissemination activities** – Consistent with OMB Circular A-130, agencies must consider, at each stage of the information life cycle, the effects of decisions and actions on other stages of the life cycle. Accordingly, to the extent permitted by law, agencies must design new information collection and creation efforts so that the information collected or created supports downstream interoperability between information systems and dissemination of information to the public, as appropriate, without the need for costly retrofitting. This includes consideration and consultation of key target audiences for the information when determining format, frequency of update, and other information management decisions. Specifically, agencies must incorporate the following requirements into future information collection and creation efforts:
 - a. **Use machine-readable and open formats**¹⁹ – Agencies must use machine-readable and open formats for information as it is collected or created. While information should be collected electronically by default, machine-readable and open formats must be used in conjunction with both electronic and telephone or paper-based information collection efforts. Additionally, in consultation with the best practices found in Project Open Data and to the extent permitted by law, agencies should prioritize the use of open formats that are non-proprietary, publicly available, and that place no restrictions upon their use.

¹⁹ The requirements of this subsection build upon existing requirements in OMB Statistical Policy Directives No. 1 and No. 2, available at <http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/directive1.pdf> and <http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/directive2.pdf>.

- b. **Use data standards** – Consistent with existing policies relating to Federal agencies’ use of standards²⁰ for information as it is collected or created, agencies must use standards in order to promote data interoperability and openness.
 - c. **Ensure information stewardship through the use of open licenses** – Agencies must apply open licenses, in consultation with the best practices found in Project Open Data, to information as it is collected or created so that if data are made public there are no restrictions on copying, publishing, distributing, transmitting, adapting, or otherwise using the information for non-commercial or for commercial purposes.²¹ When information is acquired or accessed by an agency through performance of a contract, appropriate existing clauses²² shall be utilized to meet these objectives while recognizing that contractors may have proprietary interests in such information, and that protection of such information may be necessary to encourage qualified contractors to participate in and apply innovative concepts to government programs.
 - d. **Use common core and extensible metadata** – Agencies must describe information using common core metadata, in consultation with the best practices found in Project Open Data, as it is collected and created. Metadata should also include information about origin, linked data, geographic location, time series continuations, data quality, and other relevant indices that reveal relationships between datasets and allow the public to determine the fitness of the data source. Agencies may expand upon the basic common metadata based on standards, specifications, or formats developed within different communities (e.g., financial, health, geospatial, law enforcement). Groups that develop and promulgate these metadata specifications must review them for compliance with the common core metadata standard, specifications, and formats.
2. **Build information systems to support interoperability and information accessibility** – Through their acquisition and technology management processes, agencies must build or modernize information systems in a way that maximizes interoperability and information accessibility, to the extent practicable and permitted by law. To this end, agencies should leverage existing Federal IT guidance, such as the *Common Approach to Federal Enterprise Architecture*,²³ when designing information systems. Agencies must exercise forethought when architecting, building, or substantially modifying an information system to facilitate public distribution, where appropriate. In addition, the agency’s CIO must validate that the following minimum requirements have been incorporated into acquisition planning documents and technical design for all new information systems and those preparing for modernization, as appropriate:
- a. The system design must be scalable, flexible, and facilitate extraction of data in multiple formats and for a range of uses as internal and external needs change, including potential uses not accounted for in the original design. In general, this will involve the use of standards and specifications in the system design that promote industry best practices for information

²⁰ See OMB Circular A-119, available at http://www.whitehouse.gov/omb/circulars_a119, and OMB Memorandum M-12-08, *Principles for Federal Engagement in Standards Activities to Address National Priorities* (Jan 27, 2012), available at <http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-08.pdf>

²¹ If a data user augments or alters original information that is attributed to the Federal Government, the user is responsible for making clear the source and nature of that augmentation.

²² See Federal Acquisition Regulation (FAR) Subpart 27.4—Rights in Data and Copyrights, available at https://acquisition.gov/far/current/html/Subpart%2027_4.html

²³ Office of Management and Budget, *Common Approach to Federal Enterprise Architecture*, available at http://www.whitehouse.gov/sites/default/files/omb/assets/egov_docs/common_approach_to_federal_ea.pdf

sharing, and separation of data from the application layer to maximize data reuse opportunities and incorporation of future application or technology capabilities, in consultation with the best practices found in Project Open Data;

- b. All data outputs associated with the system must meet the requirements described in part III, sections 1.a-e of this Memorandum and be accounted for in the data inventory described in part III section 3.a; and
- c. Data schema and dictionaries have been documented and shared with internal partners and the public, as applicable.

3. Strengthen data management and release practices – To ensure that agency data assets are managed and maintained throughout their life cycle, agencies must adopt effective data asset portfolio management approaches. Within six (6) months of the date of this Memorandum, agencies and interagency groups must review and, where appropriate, revise existing policies and procedures to strengthen their data management and release practices to ensure consistency with the requirements in this Memorandum, and take the following actions:

- a. **Create and maintain an enterprise data inventory** – Agencies must update their inventory of agency information resources (as required by OMB Circular A-130)²⁴ to include an enterprise data inventory, if it does not already exist, that accounts for datasets used in the agency’s information systems. The inventory will be built out over time, with the ultimate goal of including all agency datasets, to the extent practicable. The inventory will indicate, as appropriate, if the agency has determined that the individual datasets may be made publicly available (i.e., release is permitted by law, subject to all privacy, confidentiality, security, and other valid requirements) and whether they are currently available to the public. The Senior Agency Official for Records Management should be consulted on integration with the records management process. Agencies should use the Data Reference Model from the Federal Enterprise Architecture²⁵ to help create and maintain their inventory. Agencies must describe datasets within the inventory using the common core and extensible metadata (see part III, section 1.e).
- b. **Create and maintain a public data listing** – Any datasets in the agency’s enterprise data inventory that can be made publicly available must be listed at [www.\[agency\].gov/data](http://www.[agency].gov/data) in a human- and machine-readable format that enables automatic aggregation by Data.gov and other services (known as “harvestable files”), to the extent practicable. This should include datasets that can be made publicly available but have not yet been released. This public data listing should also include, to the extent permitted by law and existing terms and conditions, datasets that were produced through agency-funded grants, contracts, and cooperative agreements (excluding any data submitted primarily for the purpose of contract monitoring and administration), and, where feasible, be accompanied by standard citation information, preferably in the form of a persistent identifier. The public data listing will be built out over time, with the ultimate goal of including all agency datasets that can be made publicly available. See Project Open Data for best practices, tools, and schema to implement the public data listing and harvestable files.

²⁴ See OMB Circular A-130, section 8(b)(2)(a).

²⁵ Office of Management and Budget, Federal Enterprise Architecture (FEA) Reference Models, available at <http://www.whitehouse.gov/omb/e-gov/fea>

- c. **Create a process to engage with customers to help facilitate and prioritize data release** – Agencies must create a process to engage with customers, through their [www.\[agency\].gov/data](http://www.[agency].gov/data) pages and other necessary means, to solicit help in prioritizing the release of datasets and determining the most usable and appropriate formats for release.²⁶ Agencies should make data available in multiple formats according to their customer needs. For example, high-volume datasets of interest to developers should be released using bulk downloads as well as Application Programming Interfaces (APIs). In addition, customer engagement efforts should help agencies prioritize efforts to improve the discoverability and usability of datasets that have already been released to the public but are not yet fully “open” (e.g., they are only available in closed, inaccessible formats). See Project Open Data for best practices and tools that can be used to implement customer engagement efforts.
- d. **Clarify roles and responsibilities for promoting efficient and effective data release practices** – Agencies must ensure that roles and responsibilities are clearly designated for the promotion of efficient and effective data release practices across the agency, and that proper authorities have been granted to execute on related responsibilities, including:
 - i. Communicating the strategic value of open data to internal stakeholders and the public;
 - ii. Ensuring that data released to the public are open (as defined in part I), as appropriate, and a point of contact is designated to assist open data use and to respond to complaints about adherence to open data requirements;
 - iii. Engaging entrepreneurs and innovators in the private and nonprofit sectors to encourage and facilitate the use of agency data to build applications and services;
 - iv. Working with agency components to scale best practices from bureaus and offices that excel in open data practices across the enterprise;
 - v. Working with the agency’s Senior Agency Official for Privacy (SAOP) or other relevant officials to ensure that privacy and confidentiality are fully protected; and
 - vi. Working with the Chief Information Security Officer (CISO) and mission owners to assess overall organizational risk, based on the impact of releasing potentially sensitive data, and make a risk-based determination.

- 4. **Strengthen measures to ensure that privacy and confidentiality are fully protected and that data are properly secured** – Agencies must incorporate privacy analyses into each stage of the information’s life cycle. In particular, agencies must review the information collected or created for valid restrictions to release to determine whether it can be made publicly available, consistent with the *Open Government Directive’s* presumption in favor of openness, and to the extent permitted by law and subject to privacy, confidentiality pledge, security, trade secret, contractual, or other valid restrictions to release. If the agency determines that information should not be made publicly available on one of these grounds, the agency must document this determination in consultation with its Office of General Counsel or equivalent.

As agencies consider whether or not information may be disclosed, they must also account for the “mosaic effect” of data aggregation. Agencies should note that the mosaic effect demands a risk-based

²⁶ OMB Statistical Policy Directives 3 and 4 describe the schedule and manner in which data produced by the principal statistical agencies will be released. Statistical Policy Directive No. 4: Release and Dissemination of Statistical Products Produced by Federal Statistical Agencies, *available at* http://www.whitehouse.gov/sites/default/files/omb/assets/omb/fedreg/2008/030708_directive-4.pdf; Statistical Policy Directive 3: Compilation, Release, and Evaluation of Principal Federal Economic Indicators, *available at* http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/statpolicy/dir_3_fr_09251985.pdf

analysis, often utilizing statistical methods whose parameters can change over time, depending on the nature of the information, the availability of other information, and the technology in place that could facilitate the process of identification. Because of the complexity of this analysis and the scope of data involved, agencies may choose to take advantage of entities in the Executive Branch that may have relevant expertise, including the staff of Data.gov. Ultimately, it is the responsibility of each agency to perform the necessary analysis and comply with all applicable laws, regulations, and policies. In some cases, this assessment may affect the amount, type, form, and detail of data released by agencies.

As OMB has noted, “The individual’s right to privacy must be protected in Federal Government information activities involving personal information.”²⁷ As agencies consider security-related restrictions to release, they should focus on information confidentiality, integrity, and availability as part of the agency’s overall risk management framework. They are required to incorporate the National Institute of Standards and Technology (NIST) Federal Information Processing Standard (FIPS) Publication 199 “Standards for Security Categorization of Federal Information and Information Systems,” which includes guidance and definitions for confidentiality, integrity, and availability.²⁸ Agencies should also consult with the Controlled Unclassified Information (CUI) program to ensure compliance with CUI requirements,²⁹ the National Strategy for Information Sharing and Safeguarding,³⁰ and the best practices found in Project Open Data. In addition to complying with the Privacy Act of 1974, the E-Government Act of 2002, FISMA, and CIPSEA, agencies should implement information policies based upon Fair Information Practice Principles and NIST guidance on Security and Privacy Controls for Federal Information Systems and Organizations.³¹ For example, agencies must:

- a. Collect or create only that information necessary for the proper performance of agency functions and which has practical utility;³²
- b. Limit the collection or creation of information which identifies individuals to that which is legally authorized and necessary for the proper performance of agency functions;³³
- c. Limit the sharing of information that identifies individuals or contains proprietary information to that which is legally authorized, and impose appropriate conditions on use where a continuing obligation to ensure the confidentiality of the information exists;³⁴
- d. Ensure that information is protected commensurate with the risk and magnitude of the harm that would result from the loss, misuse, or unauthorized access to or modification of such information;³⁵ and
- e. Take into account other publicly available information when determining whether particular information should be considered PII (as defined in part I of this Memorandum).

²⁷ See OMB Circular A-130, available at http://www.whitehouse.gov/omb/Circulars_a130_a130trans4/

²⁸ NIST FIPS Publication 199 “Standards for Security Categorization of Federal Information and Information Systems”, available at <http://csrc.nist.gov/publications/fips/fips199/FIPS-PUB-199-final.pdf>

²⁹ Executive Order 13556, Controlled Unclassified Information, available at <http://www.whitehouse.gov/the-press-office/2010/11/04/executive-order-13556-controlled-unclassified-information>.

³⁰ The White House, *National Strategy for Information Sharing and Safeguarding* (December 2011), available at <http://www.whitehouse.gov/the-press-office/2012/12/19/national-strategy-information-sharing-and-safeguarding>

³¹ See NIST Special Publication 800-53 “Security and Privacy Controls for Federal Information Systems and Organizations”, available at <http://csrc.nist.gov/publications/drafts/800-53-rev4/sp800-53-rev4-ipd.pdf>

³² See OMB Circular A-130, section 8(a)(2).

³³ See OMB Circular A-130, section 8(a)(9)(b).

³⁴ See OMB Circular A-130, section 8(a)(9)(c).

³⁵ See OMB Circular A-130, section 8(a)(9)(a).

5. **Incorporate new interoperability and openness requirements into core agency processes** – Consistent with 44 U.S.C. 3506 (b)(2), agencies must develop and maintain an Information Resource Management (IRM) Strategic Plan. IRM Strategic Plans should align with the agency’s Strategic Plan (as required by OMB Circular A-11),³⁶ support the attainment of agency priority goals as mandated by the Government Performance and Results Modernization Act of 2010,³⁷ provide a description of how IRM activities help accomplish agency missions, and ensure that IRM decisions are integrated with organizational planning, budget, procurement, financial management, human resources management, and program decisions. As part of the annual PortfolioStat process,³⁸ agencies must update their IRM Strategic Plans to describe how they are meeting new and existing information life cycle management requirements. Specifically, agencies must describe how they have institutionalized and operationalized the interoperability and openness requirements in this Memorandum into their core processes across all applicable agency programs and stakeholders.

IV. **Implementation:**

As agencies take steps to meet the requirements in this Memorandum, it is important to work strategically and prioritize those elements that can be addressed immediately, support mission-critical objectives, and result in more efficient use of taxpayer dollars.

Agencies should consider the following as they implement the requirements of this Memorandum:

1. **Roles and Responsibilities** – The Clinger-Cohen Act of 1996 assigns agency CIOs statutory responsibility for promoting the effective and efficient design and operation of all major IRM processes within their agency. Accordingly, agency heads must ensure that CIOs are positioned with the responsibility and authority to implement the requirements of this Memorandum in coordination with the agency’s Chief Acquisition Officer, Chief Financial Officer, Chief Technology Officer, Senior Agency Official for Geospatial Information, Senior Agency Official for Privacy (SAOP), Chief Information Security Officer (CISO), Senior Agency Official for Records Management, and Chief Freedom of Information Act (FOIA) Officer. The CIO should also work with the agency’s public affairs staff, open government staff, web manager or digital strategist, program owners and other leadership, as applicable.

A key component of agencies’ management of information resources involves working closely with the agency’s SAOP and other relevant officials to ensure that each stage of the planning process includes a full analysis of privacy, confidentiality, and security issues. Agency heads must also ensure that privacy and security officials are positioned with the authority to identify information that may require additional protection and agency activities that may require additional safeguards. Consistent with OMB Memorandum M-05-08,³⁹ each agency’s SAOP must take on a central planning and policy-making role in all agency information management activities, beginning at the earliest stages of planning and continuing throughout the life cycle of the information. In addition, if an agency’s SAOP is not positioned within the office of the CIO, the agency should designate an official within the office of the CIO to serve as a liaison to help coordinate with the agency’s privacy office.

³⁶ OMB Circular A-11, available at http://www.whitehouse.gov/omb/circulars_a11_current_year_a11_toc

³⁷ Pub. L. No. 111-352 (2011) (codified as 31 USC § 1120 note).

³⁸ In March 2012 OMB established PortfolioStat accountability sessions, engaging directly with agency leadership to assess the maturity and effectiveness of current IT management practices and address management opportunities and challenges. For FY13 OMB PortfolioStat guidance, see OMB Memorandum M-13-09, *Fiscal Year 2013 PortfolioStat Guidance: Strengthening Federal IT Portfolio Management* (Mar. 27, 2013), available at <http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-09.pdf>.

³⁹ OMB Memorandum M-05-08, *Designation of Senior Agency Officials for Privacy* (Feb. 11, 2005), available at <http://m.whitehouse.gov/sites/default/files/omb/assets/omb/memoranda/ly2005/m05-08.pdf>

2. **Government-wide Coordination** – The Federal CIO will work with the United States Chief Technology Officer (CTO) and the Administrator of the OMB Office of Information and Regulatory Affairs (OIRA) to help improve the interoperability and openness of government information. To this end, the Federal CIO will work to establish an interagency working group supported by the Federal CIO Council. The working group should focus on leveraging government-wide communities of practice to help with the development of tools that support information interoperability and openness through repositories such as Project Open Data. Part of this work should be to share best practices related to interoperability and openness within government (e.g., Federal, state, local, and tribal). These collaborations shall be subject to statutory limitations and conducted in a way that fully protects privacy, confidentiality, confidential business information, and intellectual property rights.
3. **Resources** – Policy implementation may require upfront investments depending on the maturity of existing information life cycle management processes at individual agencies. Agencies are encouraged to evaluate current processes and identify implementation opportunities that may result in more efficient use of taxpayer dollars. However, effective implementation should result in downstream cost savings for the enterprise through increased interoperability and accessibility of the agency’s information resources. Therefore, these potential upfront investments should be considered in the context of their future benefits and be funded appropriately through the agency’s capital planning and budget processes. Some of the requirements in this Memorandum may require additional tools and resources. Agencies should make progress commensurate with available tools and resources.

In addition, tools, best practices, and schema to help agencies implement the requirements of this Memorandum can be found through the Digital Services Innovation Center and in Project Open Data.

4. **Accountability Mechanisms** – Progress on agency implementation of the actions required in this Memorandum will be primarily assessed by OMB and the public through analysis of the agency’s updates to IRM plans (part III, section 5), the completeness of the enterprise data inventory (part III, section 3.a), and the data made available in the agency’s public data listing (part III, section 3.b).

Nothing in this Memorandum shall be construed to affect existing requirements for review and clearance of pre-decisional information by OMB relating to legislative, budgetary, administrative, and regulatory materials. Moreover, nothing in this Memorandum shall be construed to reduce the protection of information whose release would threaten national security, invade personal privacy, breach confidentiality or contractual terms, violate the Trade Secrets Act,⁴⁰ violate other statutory confidentiality requirements,⁴¹ or damage other compelling interests. This Memorandum is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity by any party against the United States, its departments, agencies, or entities, its officers, employees, or agents, or any other person.

⁴⁰ 18 USC § 1905.


⁴¹ See 13 U.S.C. §§ 8, 9 and 301(g) and 22 U.S.C. § 3104.

APPENDIX G - Increasing Access to the Results of Federally Funded Scientific Research

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren 
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather data underpins the forecasting industry, and making genome sequences publicly available has spawned many biotechnology innovations. In addition, wider availability of peer-reviewed publications and scientific data in digital formats will create innovative economic markets for services related to curation, preservation, analysis, and visualization. Policies that mobilize these publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the Federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.

The Administration also recognizes that publishers provide valuable services, including the coordination of peer review, that are essential for ensuring the high quality and integrity of many scholarly publications. It is critical that these services continue to be made available. It is also important that Federal policy not adversely affect opportunities for researchers who are not funded by the Federal Government to disseminate any analysis or results of their research.

To achieve the Administration's commitment to increase access to federally funded published research and digital scientific data, Federal agencies investing in research and development must have clear and coordinated policies for increasing such access.

2. Agency Public Access Plan

The Office of Science and Technology Policy (OSTP) hereby directs each Federal agency with over \$100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government. This includes any results published in peer-reviewed scholarly publications that are based on research that directly arises from Federal funds, as defined in relevant OMB circulars (e.g., A-21 and A-11). It is preferred that agencies work together, where appropriate, to develop these plans.

Each agency plan must be consistent with the objectives set out in this memorandum. These objectives were developed with input from the National Science and Technology Council and public consultation in compliance with the America COMPETES Reauthorization Act of 2010 (P.L. 111-358).

Further, each agency plan for both scientific publications and digital scientific data must contain the following elements:

- a) a strategy for leveraging existing archives, where appropriate, and fostering public-private partnerships with scientific journals relevant to the agency's research;
- b) a strategy for improving the public's ability to locate and access digital data resulting from federally funded scientific research;
- c) an approach for optimizing search, archival, and dissemination features that encourages innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research;
- d) a plan for notifying awardees and other federally funded scientific researchers of their obligations (e.g., through guidance, conditions of awards, and/or regulatory changes);
- e) an agency strategy for measuring and, as necessary, enforcing compliance with its plan;
- f) identification of resources within the existing agency budget to implement the plan;
- g) a timeline for implementation; and
- h) identification of any special circumstances that prevent the agency from meeting any of the objectives set out in this memorandum, in whole or in part.

Each agency shall submit its draft plan to OSTP within six months of publication of this memorandum. OSTP, in coordination with the Office of Management and Budget (OMB), will review the draft agency plans and provide guidance to facilitate the development of final plans that are consistent with the objectives of this memorandum and, where possible, compatible with the plans of other Federal agencies subject to this memorandum. In devising its final plan, each

agency should use a transparent process for soliciting views from stakeholders, including federally funded researchers, universities, libraries, publishers, users of federally funded research results, and civil society groups, and take such views into account.

3. Objectives for Public Access to Scientific Publications

To the extent feasible and consistent with law; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, the results of unclassified research that are published in peer-reviewed publications directly arising from Federal funding should be stored for long-term preservation and publicly accessible to search, retrieve, and analyze in ways that maximize the impact and accountability of the Federal research investment.

In developing their public access plans, agencies shall seek to put in place policies that enhance innovation and competitiveness by maximizing the potential to create new business opportunities and are otherwise consistent with the principles articulated in section 1.

Agency plans must also describe, to the extent feasible, procedures the agency will take to help prevent the unauthorized mass redistribution of scholarly publications.

Further, each agency plan shall:

- a) Ensure that the public can read, download, and analyze in digital form final peer-reviewed manuscripts or final published documents within a timeframe that is appropriate for each type of research conducted or sponsored by the agency. Specifically, each agency:
 - i) shall use a twelve-month post-publication embargo period as a guideline for making research papers publicly available; however, an agency may tailor its plan as necessary to address the objectives articulated in this memorandum, as well as the challenges and public interests that are unique to each field and mission combination, and
 - ii) shall also provide a mechanism for stakeholders to petition for changing the embargo period for a specific field by presenting evidence demonstrating that the plan would be inconsistent with the objectives articulated in this memorandum;
- b) Facilitate easy public search, analysis of, and access to peer-reviewed scholarly publications directly arising from research funded by the Federal Government;
- c) Ensure full public access to publications' metadata without charge upon first publication in a data format that ensures interoperability with current and future search technology. Where possible, the metadata should provide a link to the location where the full text and associated supplemental materials will be made available after the embargo period;

- d) Encourage public-private collaboration to:
 - i) maximize the potential for interoperability between public and private platforms and creative reuse to enhance value to all stakeholders,
 - ii) avoid unnecessary duplication of existing mechanisms,
 - iii) maximize the impact of the Federal research investment, and
 - iv) otherwise assist with implementation of the agency plan;
- e) Ensure that attribution to authors, journals, and original publishers is maintained; and
- f) Ensure that publications and metadata are stored in an archival solution that:
 - i) provides for long-term preservation and access to the content without charge,
 - ii) uses standards, widely available and, to the extent possible, nonproprietary archival formats for text and associated content (e.g., images, video, supporting data),
 - iii) provides access for persons with disabilities consistent with Section 508 of the Rehabilitation Act of 1973,¹ and
 - iv) enables integration and interoperability with other Federal public access archival solutions and other appropriate archives.

Repositories could be maintained by the Federal agency funding the research, through an arrangement with other Federal agencies, or through other parties working in partnership with the agency including, but not limited to, scholarly and professional associations, publishers and libraries.

4. Objectives for Public Access to Scientific Data in Digital Formats

To the extent feasible and consistent with applicable law and policy²; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, digitally formatted scientific data resulting from unclassified research supported wholly or in part

¹ Section 508 Of The Rehabilitation Act, as amended, available at: <https://www.section508.gov/index.cfm?fuseAction=1998Amend>

² These policies include, but are not limited to OMB Circular A-130, Management of Federal Information Resources, available at: http://www.whitehouse.gov/omb/circulars_a130_a130trans4

by Federal funding should be stored and publicly accessible to search, retrieve, and analyze. For purposes of this memorandum, data is defined, consistent with OMB circular A-110, as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens. Each agency's public access plan shall:

- a) Maximize access, by the general public and without charge, to digitally formatted scientific data created with Federal funds, while:
 - i) protecting confidentiality and personal privacy,
 - ii) recognizing proprietary interests, business confidential information, and intellectual property rights and avoiding significant negative impact on intellectual property rights, innovation, and U.S. competitiveness, and
 - iii) preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden;
- b) Ensure that all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long-term preservation and access cannot be justified;
- c) Allow the inclusion of appropriate costs for data management and access in proposals for Federal funding for scientific research;
- d) Ensure appropriate evaluation of the merits of submitted data management plans;
- e) Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies;
- f) Promote the deposit of data in publicly accessible databases, where appropriate and available;
- g) Encourage cooperation with the private sector to improve data access and compatibility, including through the formation of public-private partnerships with foundations and other research funding organizations;
- h) Develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan;

- i) In coordination with other agencies and the private sector, support training, education, and workforce development related to scientific data management, analysis, storage, preservation, and stewardship; and
- j) Provide for the assessment of long-term needs for the preservation of scientific data in fields that the agency supports and outline options for developing and sustaining repositories for scientific data in digital formats, taking into account the efforts of public and private sector entities.

5. Implementation of Public Access Plans

Some Federal agencies already have policies that partially meet the requirements of this memo. Those agencies should adapt those policies, as necessary, to fully meet the requirements. Once finalized, each agency should post its public access plan on its Open Government website.

The agency plan shall not apply to manuscripts submitted for publication prior to the plan's effective date or to digital data generated prior to the plan's effective date. The effective dates can be no sooner than the publication date of the agency's final plan.

OSTP will oversee implementation through regular meetings with agencies. Each agency shall provide updates on implementation to the Directors of OSTP and OMB twice yearly; these updates shall be submitted by January 1 and July 1 of each year for two years after the effective date of the agency's final plan. An agency may amend its public access plan consistent with these objectives, in consultation with OSTP and OMB.

6. General Provisions

Nothing in this memorandum shall be construed to impair or otherwise affect authority granted by law to an executive department, agency, or the head thereof; or functions of the Director of OMB relating to budgetary, administrative, or legislative proposals.

Consistent with the America COMPETES Reauthorization Act of 2010, nothing in this memorandum, or the agency plans developed pursuant to it, shall be construed to authorize or require agencies to undermine any right under the provisions of title 17 or 35, United States Code, or to violate the international obligations of the United States. This memorandum is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity, by any party against the United States; its departments, agencies; or entities, its officers, employees, or agents; or any other person.