Colorado State University

ISTeC Data Management Committee

Committee Findings and Recommendation

January 31, 2014

Executive Summary

Data management is a core competency in most research projects. As datasets grow in volume, velocity and variety, they can quickly become unwieldy. The CSU research community, like those at many other research institutions, is rapidly becoming overwhelmed with data management, diverting time and energy from research projects.

A recent survey shows CSU's research community has challenges in storage availability, organization, preservation and collaboration. While there is no panacea for these challenges, much can be done to reduce or eliminate hurdles in the research function. Four initiatives can begin to ease these challenges at CSU:

1. Learn from one another through Affinity groups. Like assembling a puzzle, individuals may hold the key to certain pieces of the puzzle and, through the sharing of ideas and knowledge, puzzles may be partially or wholly assembled.
2. Spread knowledge broadly through easily accessible vignettes that focus on specific data management challenges faced by researchers.
3. Build an information technology infrastructure, including both storage and compute capacity, to facilitate research.
4. Merge the research process with data curation practices to make data management seamless for the CSU research community.

Confronting this issue now is important, as some research is churning in the data management morass. Other research institutions are further along than CSU in addressing data management challenges. To remain a viable, competitive and thriving research institution, CSU needs to take action. It is recommended that CSU create a work plan, based on these initiatives, within the next 120 days with a plan to execute the work plan in fiscal 2015.

CSU's data ecosystem is complex and vital to the future success of the institution. Data management for research is a core competency CSU needs to master.

Background

The ability to generate and capture scientific data continues to accelerate. Analyzing "Big Data" promises to reveal many mysteries if researchers can figure out how to harness it. Big Data are generated by many CSU investigators, as well as through collaborations. As the amount of research data grows, managing, mining and analyzing data becomes increasingly difficult, particularly for investigators that do not have strong computer and computational skills.

Recognizing that Big Data management is an institutional challenge, CSU's Information Science and Technology Center (ISTeC) sponsored a "Big Data" Research Forum in mid-April, 2013. The Forum was well attended, with over 100 participants and focused on the challenges and opportunities facing the research community.

Recognition of the diverse and significant Big Data needs of CSU's research community resulted in the formation of the ad-hoc ISTeC Data Management Committee.

ISTeC Data Management Committee Charter

CSU's Chief Information Officer, Pat Burns, charged the ISTeC Data Management Committee "to establish strategies and directions for an institutional approach to data management." The formal charter is attached in Appendix A. The Committee elaborated on this charge by defining the Research Data Ecosystem as the organization, administration and governance of large structured and unstructured data sets created at CSU or obtained through collaboration with CSU investigators while adhering to the institution's goals, objectives and regulatory requirements.

Components agreed upon to be investigated included:

1. Organization and documentation, with a focus on **searchability** and **findability**;

2. Administration, with attention to security, accessibility and preservation;

3. Infrastructure, specifically data storage and transport vehicles; and
4. Governance, to assure practices that comply with requirements of the funding entities and university policies.

Committee Members

The Committee members contributing to this recommendation are:
Mr. Scott Baily, Academic Computing & Network Services
Dr. Patrick Burns, Dean of Libraries & CIO
Dr. Richard Casey, IDRC & HPC
Ms. Nancy Hunter, Libraries
Dr. Andrew Jones, Cooperative Institute for Research in the Atmosphere
Dr. Nicole Kaplan, Natural Resources Ecology Laboratory
Dr. Rick Lyons, Infectious Disease Research Center
Mr. Scott Novogoratz, College of Veterinary Medicine & Biomedical Sciences
Mr. Ed Peyronnin, College of Agricultural Sciences
Dr. Richard Slayden, College of Veterinary Medicine & Biomedical Sciences
Mr. Shea Swauger, Libraries

Big Data Landscape

What is Big Data? A common definition is based not on the absolute size of a particular dataset, but a "dataset whose size is beyond the ability of readily available tools to capture, store, manage and analyze[1]". Big data comes from many sources including transactional data, devices and machines. In general, there are certain characteristics ascribed to Big Data, often referred to as the "V's" which include:

[1] From McKinsey Global Institute, Big data: The next frontier for innovation, competition and productivity

1. Volume – amount of data to manage
2. Velocity – speed at which the data presents itself
3. Variety – data that defies traditional database models

While many strategies are evolving for Big Data management, there is no panacea.

CSU is not alone among research institutions in its quest to find better research data management solutions. For example, Purdue University has identified "Big Data" as a major thrust for inter-disciplinary research and education and is hiring faculty positions within its computer science department to fill the void. Purdue's focus includes data infrastructure (storage and networking), system software infrastructure (OS, programming models), scalable analytics (data mining and machine learning at scale), analytics presentation (information visualization) and data-enabled methods (data curation and preservation) and applications in domain sciences.

The corporate world also sees Big Data as an opportunity to gain competitive advantage. Strategy consultants MGI and McKinsey are doing significant work in the Big Data domain[2] and cite Big Data as the "next frontier for innovation, competition and productivity."

According to a recent study, the United States government believes it can save 14% of its annual budget, or $500B annually, by leveraging Big Data[3]. This estimate is based on increases in productivity/efficiency, enhanced security and better ability to predict trends; all factors in permitting the government to work smarter.

[2] See study at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
[3] See study at http://www.meritalk.com/pdfs/emc-big-data/Smarter_Uncle_Sam_Infographic.pdf

CSU Big Data Activities and Groups

There are various activities in research, education and practice that address the growing needs for managing Big Data. The groups taking leadership in these activities represent diversity in data management needs, applications, and cyber-infrastructure development. Some examples, which are familiar to the few members of the IsTec Data Management Committee include:

CSU's Libraries convened a faculty librarian research team to investigate researchers' data curation practices and needs in mid-2012. While the results of this study are not yet available publicly, themes from the focus groups participating in this study recognized that many researchers are struggling with data management and would welcome assistance[4].

CSU's Computer Science department now has a faculty member, Sangmi Lee Pallickara, whose research is focused on Big Data for the sciences, specifically, issues related to storage, analytics, integration, metadata, and visualization. Dr. Pallickara has major research funding from the Department of Homeland Security to study the Big Data Analytics of Epidemic Outbreaks. She also teaches a graduate level course in Big Data, CS581.

CSU's College of Business offers several Business Intelligence courses focusing on Big Data[5] as well as several courses with Big Data occupying a portion of class time[6] Stephen Hayne, CSU Professor of Computer Information Systems, also directs the Computer Information Systems Center for Business Analytics (CISCBA) within the College

CitSci.org is directed by Dr. Greg Newman at the Natural Resources Ecology Lab, (**NREL**), as an eco-informatics oriented data management platform designed to promote and facilitate citizen involvement in scientific research. Citizen Science

[4] See McLure, M., Level, A. V., Cranston, C. L., Oehlerts, B., & Culbertson, M. (in press). Data curation: A study of researcher practices and needs. *portal: Libraries and the Academy*.

[5] Business Intelligence (CIS570), Applied Data Mining and Business Analytics (CIS 575), Business Database Systems (CIS655), and Data Visualization (CIS690C)

[6] CIS200 (1 week), BUS630 (1 week), CIS600 (2 weeks), BUS690 (1 week), CIS570 (several weeks), CIS575 (several weeks), CIS655 (several weeks), BUS690C (several weeks)

empowers individuals to pursue their interests in the scientific world through developing and implementing projects following the scientific method. Community members are trained as volunteers, contribute to investigations of scientific questions for existing projects and new projects are being created regularly! Thus far, CitSci.org supports over 78 citizen science projects across the world and contains and curates over 46,000 data records. The CitSci.org system stands out for its customization features that allow you to include your creativity into your research and inter-operability functions that facilitate data sharing between different organizations.

The following three groups lead activities and have been identified by the ISTeC Research Committee as being established, but with demand to grow, and of interest to different research domains through converging themes and/or methods. These groups may be well positioned to serve as Affinity Groups for exploring and expanding data management capacity on-campus (see Initiative 1 in recommendations below):

> CSU's GeoSpatial Centroid is dedicated to facilitating campus-wide communication about events, research, and education in geospatial sciences. The Centroid is a resource and research center at Colorado State University established to provide students, faculty, and the Colorado community with information about GIS at CSU and how these activities link to broader statewide, regional and global initiatives[7].
>
> CSU supports a joint effort between the University Libraries and the Colorado Water Institute, to curate a Water Resources Archive and create a Water Center to facilitate interdisciplinary research through funding opportunities and access to data, information, expertise and tools[8]. http://www.cwi.colostate.edu/
>
> CSU's Next Generation Sequencing Core was created to to provide campus-wide DNA sequencing services to faculty, graduate students, post-doctoral students, and other researchers, and to offer those services to external government, non-profit, and commercial organizations.

[7] http://gis.colostate.edu/
[8] http://www.cwi.colostate.edu/

Data, the lifeblood of organizations, is increasing in volume, velocity and variety. Information technology (IT) is critical for harnessing the power of this big data to derive business value.

Committee Activities

Taking its charge, the Committee acknowledged there was a great deal of anecdotal information surrounding research data management. The Committee also recognized it was difficult to understand the breadth and depth of the data management problems facing the research community. The Committee agreed to survey the campus research community about its attitude towards data management.

Attitude Survey

The Committee assembled a survey instrument to assess which campus groups faced data management challenges and specific information about those challenges. In working with the Office of the Vice President for Research, a list of 1539 campus researchers were identified to receive the survey. Of those receiving the survey, 260 responded for a response rate of almost 17%. In responding to the survey, individuals were asked to identify their position and College affiliation. Most Colleges were well represented with over 30 responses each, with the exceptions of the Colleges of Business (3 responses) and Liberal Arts (14 responses). The survey results are included in Appendix B.

Survey Observations

By stratifying the challenges facing CSU researchers, the following observations can be made from the survey results:

Most Challenging Issues (95% CI above the middle response)

1. Funding the infrastructure (hardware, software and staffing) to support my data
2. Preserving my research data long term is a safe and secure environment
3. Organizing, cataloging, documenting (metadata) and managing data acquired for my research
4. Facilitating interactive collaboration by making research data files available to others

Moderately Challenging Issues (95% CI contains the middle response)

1. Finding and utilizing a convenient and safe medium to store data needed for my research
2. Analyzing my research data (e.g., the large amount of data makes it difficult to analyze)

Less Challenging Issues (95% CI below the middle response)

1. Providing open access to my research data to comply with sponsors' data management requirements
2. Determining which parts of my research data, if any, needs to be available to the public
3. Protecting intellectual property
4. Finding appropriate software tools to facilitate analysis of my research data
5. Moving research data to where it is needed
6. Protecting sensitive data from unauthorized access
7. Protecting CSU's and my intellectual property rights regarding my research data

Themes from the survey's verbatim responses to the question, "Are there any other activities related to research data that are challenging?" include:

- Bioinformatics is difficult
- Training needed
- Unstructured data presents new challenges
- Collaboration opportunities wanted
- Interdisciplinary approach needed for data management

- Truly understand data, information, analysis and results
- Resource availability (hardware, software, storage capacity, expertise)

From the 130 responses to the question, "What one thing would you want the ISTeC Data Management Committee to do to help you solve your Research Data challenges?" here are some representative answers:

- A cloud storage system accessible from on or off campus
- Communication to the campus community as things are constantly changing
- Data query support
- Educate research and faculty on the options for storage, management, sharing and moving large quantities of data
- Faster file transfer; e.g. for 1 TB of data
- Funding the infrastructure
- Have a dedicated data manager for my department/projects as part of my research team
- Keep track of lab note content electronically
- Increase assistance for analysis of big data
- Invest in computational power
- License software at the University level
- More powerful computers and training on Linux, python and metadata analysis
- On-line backup for large amount of data
- Provide permanent archival capabilities with an accompanying permanent web interface
- Visible contact points for assistance with big data

Based on the survey results, the Committee brainstormed possible solutions which appear in Appendix C.

Recommendation

To move CSU to a better research data management environment and to protect the institution from data mismanagement, the Committee recommends four initiatives. To carry out these initiatives, the Committee also recommends relying on the cooperative efforts of two existing campus groups, ISTeC and IAC. Collaboration between these groups will provide informed solutions to the complex data management issues facing CSU today.

The Information Science and Technology Center (ISTeC) (http://istec.colostate.edu/) is a faculty-driven university-wide organization for promoting, facilitating, and enhancing CSU's research, education, and outreach activities pertaining to the design and innovative application of computer, communication, and information systems. ISTeC encompasses Information Science & Technology (IS&T) activities among faculty throughout the various existing colleges of the university cutting across college boundaries.

The ITEC Advisory Council (IAC) (http://www.acns.colostate.edu/IAC) considers operational aspects of the IT environment from a professional IT perspective (Policy ID: 4-1018-008). Membership includes the senior IT representative from the colleges and VPs' offices and is responsible for maintaining the preponderance of the IT infrastructure on campus. IAC reviews matters of policy, strategy, and management of the IT environment and, when necessary, forwards to the IT Executive Committee (ITEC) for its consideration. ITEC is chaired by the Provost with the President's cabinet, one member from the faculty council and one from the College Administrators Council, as voting members. On operational matters where consensus is achieved, the IAC may take action unilaterally.

The two groups and the VP-IT form the campus Affinity Group Technology Resource. The organizations will choose how they wish to work together. They may wish to form a combined subcommittee or simply conduct business electronically. They will decide how to define the scope and constitution of an affinity group. If it is defined too narrowly, efforts will be duplicated for IT staffs. If it is defined too broadly, the group may become paralyzed by lack of consensus. It will also provide the tools, resource and guidance to support affinity centers as the start and evolve.

Initiative 1 – Affinity Groups

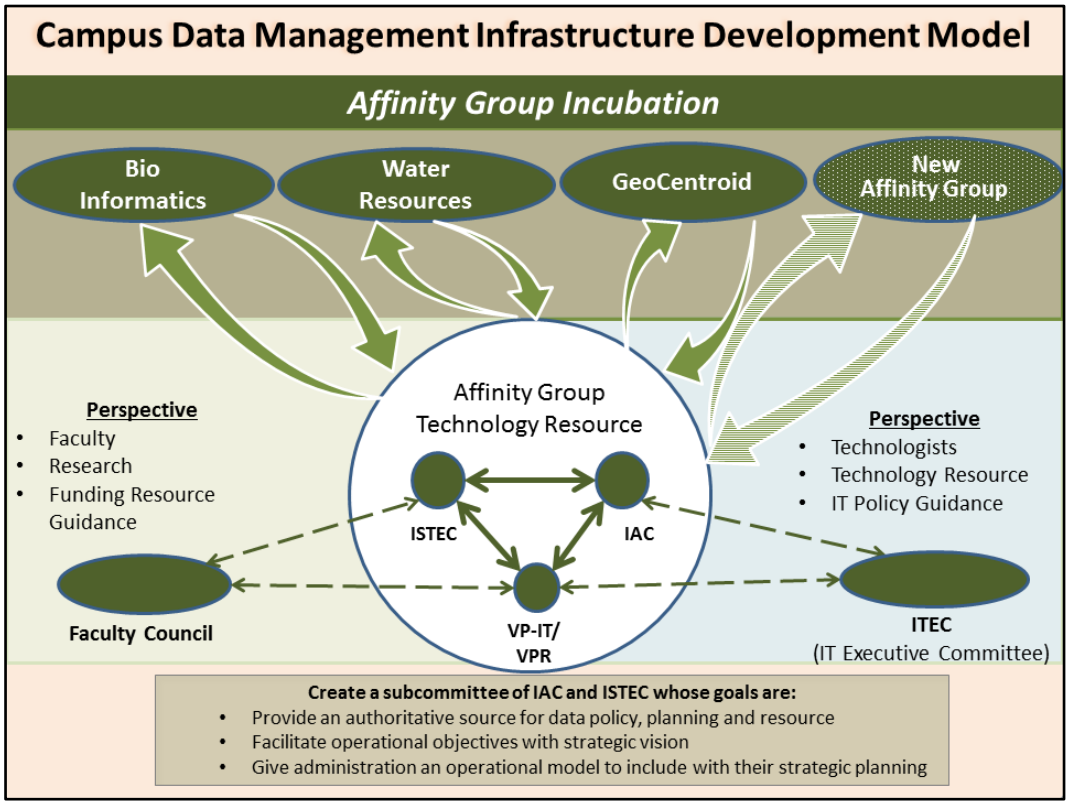
Form Affinity Groups focusing on data management challenges within the University research community.

CSU is rich with knowledge. The Committee believes that much of the knowledge needed to solve many data management challenges is already available on the CSU campus. However, CSU lacks mechanisms to share specific knowledge. For example, one Committee member, Dr. Rick Lyons, took the initiative to start a Bioinformatics Affinity Group. With a simple agenda to begin the discussion and a single email message, over 35 individuals attended the first meeting and, by consensus, agreed to continue meeting at least once per month to share information. The Committee recommends the following steps to gain traction for these Affinity Groups:

- Raise Awareness of Common Data Management Issues Confronting the CSU Research Community.
    - Raise awareness about opportunities to build Affinity Groups through a publicity campaign.
- Use crowdsourcing practices to identify topics and generate interest. Create a structure to foster the creation and sustainability of Affinity Groups (see Figure 1).
    - Leverage existing resources, such as the ISTeC Research Advisory Committee and the ITEC Advisory Council (IAC), to provide structure and sustainability for Affinity Groups.
    - Provide support for individuals to initiate campus-wise discussions on a specific research data management topic.
    - Recruit and/or elect individuals to lead Affinity Groups, as groups are created.
    - Dissolve Affinity Groups if there is no longer interest.
- Create a starter set of Affinity Groups.

- Leveraging existing initiatives to form Affinity Groups for GeoSpatial Centroid, Water Center, and Bioinformatics.
- Using information collected from the campus survey and other anecdotes, identify additional initial Affinity Groups.
- Form additional initial Affinity Groups with broad subject matter and spin off subgroups as needed.
- Begin Affinity Groups using "lightning talks" to assess and generate interest.

Figure 1



Initiative 2.A – Educational and Outreach Opportunities

2.A Create data management educational opportunities for formal training among data practitioners within the CSU research community.

As an institution with both an education and research mission, CSU is well equipped to train the research community in contemporary data management techniques. Data management has become a required step in conducting research and CSU should create opportunities to reach out to researchers regarding the capabilities, challenges and opportunities that are related to managing data.

As data management tools and techniques continue their rapid evolution, it is important that CSU create opportunities for the research community to learn about new tools and techniques.

Data management is largely a practice and, as with most practices, there are best practice guidelines to follow. These best practices can form the governance model for CSU's research data management. By promoting best practice data management guidelines, the CSU research community be more efficient and effective, limiting data loss and the amount of time investigating various tools and methods to get things done. For example, with many tools available to move large datasets from one place to another, CSU can endorse and support tools designed to easily and efficiently transfer large data sets. Additionally, research methods and data management processes are often quite similar. By identifying and publishing local (specific to CSU) pipelines for particular scientific processes and methods; e.g., how to sequence and analyze DNA, the research community no longer needs to rely solely on informal mechanisms to discover how to get things done.

The Committee recommends the following steps to build the Data Management educational and training program for research associates and scientists in the research community:

- Identify target audiences and topics for data management education.
    -

- Use information collected from the Research Data Management survey to identify opportunities for education.
- Ask for suggestions among the research community.
- Utilize feedback from Affinity Groups and the ISTeC Research Advisory Committee.
- Set up brown bag sessions for sharing approaches and demos among practitioners on campus.

- Identify data management experts to lead the educational effort an identify opportunities for training a broader audience.
  - Identify information scientists and technology experts who are willing to "train the trainer" (e.g. lab managers supervisors, IT professionals) and/or create educational sessions and videos.
  - Draw upon the pool of CSU campus individuals who are trained in the information and/or technology sciences, as well as those who are in training.
  - Create reusable data management education snippets, shorts, and tutorials using Echo360 and/or RamCT (e.g. tutorial on CSU's Digital Repository or techniques to anonymize data.) Provide a review-process, perhaps through the ISTeC RAC, to manage quality and content accuracy prior to publishing.
  - Explore recruiting instructors, developing and co-teaching a data management course through interdisciplinary collaborations within the colleges (Data Management 101) for students.
  - Develop step-by-step tools, such as a self-deposit module for scientists submitting data into the Institutional Repository.
  - Utilize experienced educational material professionals, such as the Computer Application Training (CAT) group

housed in the Morgan Library and the campus Instructional Technology designers for guidance in instructional video preparation.

2.B Plan informal data management outreach activities for members of the research community. Bring visibility to the issues, skills and tasks involved in managing research data.

- Build the Data Management outreach program for the CSU research community:
    - Explore ways to encourage the CSU research community to educate themselves on data management when appropriate by building a library of links to data management resources, such as, how-to videos and templates, hosted by CSU (e.g. PDI, presentations) or partner institutions.
    - Find keynote speakers with key messages to gain a broader perspective on specific issues as related to broader expectation and possibilities enabled by the development of a more robust cyber-infrastructure.
    - Model presentations after the Distinguished Lecture series to ignite interest in contributions of data to science today (e.g. eScience, Bioinformatics, eco-informatics, crowd-sourcing, mash-ups).

Initiative 3 – Physical Infrastructure

Build physical IT infrastructure to meet the current and anticipated needs of the University research community.

The underpinnings of data management lie in the technology available to support it. The Committee believes certain components of CSU's current campus infrastructure, such as storage and compute limitations, inhibit research opportunities. It's often unrealistic and impractical to expect each researcher to fund the acquisition of data storage, compute capacity and analytic tools to carry out their research, as there are often just short-term needs for this technology infrastructure. While there are

several 3rd party mechanisms, such as Amazon's cloud and Microsoft Azure, available to fill this void, the Committee finds it difficult to endorse these solutions, as there are no guarantees for protection and security. Additionally, the Committee must discourage use of these 3rd party solutions as the end-user license agreements typically conflict with University and State of Colorado procurement policies as well as the guidelines of the granting entities, such as the provision that all research data must be housed within the U.S. Consequently, the Committee recommends the following enhancements to CSU infrastructure to better meet the needs of the CSU research community:

- Continue to expand cloud data storage specifically designated for CSU research, permitting the CSU research community to access their data both on and off campus, while also providing the opportunity to share datasets and collaborate with non-CSU researchers
- Enhance compute capacity to provide resources for compute-intensive research.
- Create a research network to permit research data to flow more easily throughout the CSU campus and permit the campus research network to be connected with other research networks throughout the U.S. A promising solution the Committee recommends for further investigation is Globus Online[9], specifically for managing research data.

Initiative 4 – Administrative Framework

Provide an administrative framework for managing research data.

From the survey responses and anecdotal information, it is clear to the Committee that the CSU research community is struggling with many aspects of data management including:

- Searchability and findability of prior and/or relevant research data,

[9] See link at http://www.globus.org

- Compliance with sponsors' regulatory requirements, and
- Broaden and empower large-scale utilization of an individual researcher's data, information, analysis and findings.

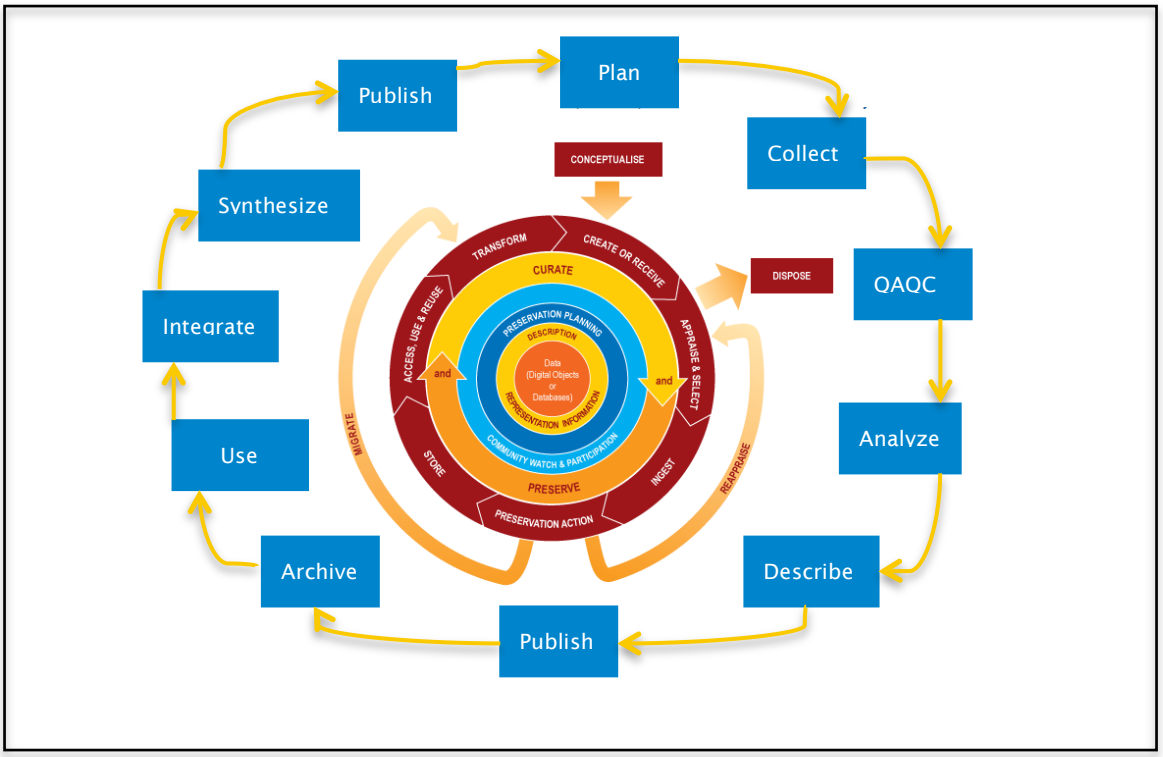
By providing a framework for campus researchers, CSU can facilitate a common data management approach for organization and documentation. The Committee recommends the conceptual Curation Life Cycle[10] model as the basis for CSU's research data management framework. Furthermore, the Committee recommends:

- Building awareness of this model,
- Identifying how and where data management practices can be incorporated,
- Creating an on-ramp as needs arise, and
- Assimilating researchers into the framework as both creators and consumers of curated data and information.

General steps observed and reported by scientists in conducting research are shown in parallel to related or similar activities included in the Data Life Cycle Model shown in Figure 2.

Figure 2

[10] See link at http://www.dcc.ac.uk/resources/curation-lifecycle-model



The Curation Lifecycle model was developed by researchers at the International Digital Curation Center and communicates several fundamental ideas that enable successful data management as it relates to preserving and curating data and information that is produced in research and scholarship.

- One of the main motivations of managing data is to enable its use and reuse.
- Data is dynamic and undergoes change over time based on the needs of the community that uses it. This can include adding descriptive information about the data that allow it to be found and understood more readily.
- Data requires repeated appraisal and curation in order to continue to meet the needs of its user community. This might mean formats might need to be migrated over time.

- Not all data should be saved and that sometimes it is appropriate to dispose of data.

The Committee recognizes an alignment between the scientific process, with familiar steps for conducting research and the Data Life Cycle, to inform the CSU Libraries Data Management strategy, policies and practices.

- Data management planning is expected in the beginning of the cycle when experiments are designed, and includes required supplemental documents with proposals submitted to NSF and other sponsors
- Data quality needs to be both assessed and assured to produce data of high integrity
- Metadata are necessary to describe methods, conditions, data and other digital objects
- Results are published and it is expected that supporting data and other information used in the analysis be accessible
- Prior research data is searchable, finable and accessible so it may be used and integrated with other data as researchers extend prior research, ask new questions and are funded for new research.
- Synthetic research produces new results for publication

Many of the steps in research relate to steps in the Data Life Cycle model, with opportunities for data managers and librarians to facilitate the research function, provide necessary data management services, and create opportunities for training.

Conclusion

The committee's charter includes the metaphor, "data ecosystem" to describe a large, dynamic, complex, self-sustaining network of people, process and technology. "Big Data" is central to CSU's research. This analysis attempts to

define and understand how to manage and preserve Big Data. The Committee concludes that CSU's data ecosystem is a shared responsibility among campus researchers, technicians and administration and no single solution exists to resolve the many issues surrounding research data management. The research community's needs are central to any course of action; however they need tools and guidance from technicians and information scientists to achieve their goals.

While the Committee cannot endorse co-opting other CSU programs, the Committee recommends leveraging existing campus resources as much as possible to carry out the four initiatives. The Committee recommends that ISTeC, IAC and the VPs for IT and Research, work in partnership to build systems, policies and practices for CSU's data ecosystem. Assuming these entities have the time and interest in moving these four initiatives forward, they can:

- Be a resource for affinity groups
- Formalize workflows to streamline the data management process
- Assist in developing funding models,
- Review, approve and implement policies,
- Establish metadata standards and
- Develop training resources.

To move from studying the problem to taking action:

1. Foster the three existing affinity groups serving as exemplars; the Bioinformatics group, Geospatial Centroid and the Colorado Water Institute. These established collaborative groups are organized around themes that represent a good cross-section of the university that can work with the IT professional staffs to identify needs and develop resources.

2. Enlist instructional technology designers to develop 3 – 5 minute online best practices training videos using the campus licensed lecture capture software.

3. Leverage the work already done with CSU's physical IT infrastructure to provide a more rational campus-wide structure for both data centers and data storage. Remove technological barriers and hurdles that prevent researchers from moving forward with their projects.

4. Merge the CSU Libraries data curation initiatives with those of the Committee to provide a seamless solution that serves both the compliance and practical data management requirements.

There is an urgent need to move forward with these initiatives, as data management issues plague many CSU researchers and divert attention from their research projects. The Committee believes CSU is in a catch-up mode and recommends that a work plan be developed within the next 120 days to begin its execution at the start of the next fiscal year, July 1, 2014. Components of the work plan are:

1. Vet, revise (if necessary) and accept the ISTeC Data Management Committee's recommendation.
2. Identify the "low-hanging" fruit where success is likely to promote visibility and gain inertia.
3. Adopt models that work elsewhere in similar circumstances.
4. Develop scope, identify resources and establish an implementation timeline for each of the four initiatives.
5. Enlist and/or form the leadership group(s) to oversee and fulfill the Committee's recommendation.
6. Identify a funding model to support the planned activities.
7. Design and execute a communication plan.

The task before us is a complex, collaborative endeavor. However, information is at the core of our mission. As campus constituents, granting agencies and/or citizens around the state use this information, we must ensure its sustained availability. Collaboration and data sharing is now required by major granting agencies. To remain competitive as a research institution, our endeavor's success is essential. The relationships we will build in the organization as a result of the proposed infrastructure will enhance our collaborations, improve our research and may serve as a model for other institutions.

Appendix A

ISTeC Data Management Committee
Charge to the Committee
Revised on July 10, 2013

The ISTeC Data Management Committee (hereinafter the 'Committee') is hereby constituted as an ad hoc committee to recommend strategies and directions for an institutional approach to Research Data Management. The initial committee members are: Scott Novogoratz, CVMBS (chair); Scott Baily, ACNS; Rick Casey, Manager of Central HPC Systems; Nancy Hunter, CSU Libraries; Andy Jones, CIRA; Nicole Kaplan NREL (SGS-LTER); Rick Lyons, IDRC (representing the Office of the VP for Research); Ed Peyronnin, College of Agricultural Sciences; and Ric Slayden, CVMBS; and. The Committee shall meet biweekly, or as often as the Chair deems necessary.

Definition of Research Data Management Ecosystem – The organization, administration and governance of large structured and unstructured data sets created or obtained through CSU research activities while adhering to the institution's goals, objectives and regulatory requirements.

Components of the Research Data ecosystem to investigate include, but are not limited to:

1. Organization (Committee can provide guidelines, many already developed, for researchers)
    a. File types, sizes, formats, standards
    b. Level of preservation support (http://lib.colostate.edu/repository/csu-digital-repository-preservation-format-support-policy)
    c. Self-Documentation (meta-data provided), 'findability,' data categorization, searchability
    d. Data associated with scholarly communications (i.e. pairing of data with research articles)
2. Administration

   a. Business model and sustainability
   b. Data Management
      i. Data security
      ii. Data access controls
      iii. Data protection and safety
      iv. Data privacy
      v. Data preservation
   c. Work Flow
      i. Integration with the research work flow
      ii. Ability to locate and retrieve specific research data quickly and accurately
      iii. Processes to store Research Data
3. Infrastructure
   a. Storage and preservation of Research Data including resource expectations from central, colleges, departments and units
   b. Transport of research data
4. Governance
   a. Legal and Compliance
      i. Intellectual property (IP) protection, copyright vs. IP issues
      ii. Compliance with CSU and granting entities' policies regarding research data
   b. Policy recommendations, if appropriate

If the committee wishes, discuss how the Committee's task overlaps, supplements, or complements Research Analytics. Should the Committee elect to address this, the Committee should have a dialogue with Laura Jensen from IR, and Jim Folkestad of the School of Education.

The Committee shall endeavor to complete an initial analysis by mid-December, 2013. Outcomes from this initial analysis should include a landscape analysis of CSU's Research Data management practices and problems, suggestions to improve the Research Data ecosystem and recommended next steps.

Appendix B

Survey Results

Appendix C

Brainstorming Ideas

Armed with the survey information the ISTeC Data Management Committee brainstormed possible solutions to the problems facing CSU researchers.

1. Help establish affinity groups around challenges
2. Offer training seminars
3. Look at models that worked at other institutions
4. Replicate models that have worked in specific disciplines; e.g., meteorology
5. Find funding sources outside CSU; e.g., NSF Office of Cyberinfrastructure
6. Assemble research scientists and research associates to identify and tackle their data management problems
7. Associate metadata with research plan to enhance repeatability with the aim of fulfilling requirements for data submission to journals and funding entities, such as NIH
8. Develop common language to discuss data management among scientists from multiple disciplines
9. Build out physical infrastructure (storage & network), including training and available expertise to help people use it
10. Apply versioning methodology for all research datasets