Vol. 4, No. 1 | December 2016
RENCI WHITE PAPER SERIES

From Data Center 1.0 to Data Center 3.0:

TRANSFORMING THE STORAGE, ACCESS, AND (RE)USE OF RESEARCH DATA FOR BETTER SCIENCE

The Team

W. CHRISTOPHER LENHARDT
Domain Scientist, Environmental Data Science and Systems, Renaissance Computing Institute (RENCI)

BRIAN BLANTON
Senior Research Scientist, Renaissance Computing Institute (RENCI)

CHARLES SCHMITT
Chief Technology Officer, Director of Informatics (RENCI)

Additional writing support from Anne Frances Johnson and Kathleen Pierce.

Summary

Scientific data centers have long played a crucial role in research and innovation. Over the past several decades, data centers have influenced, and been influenced by, the evolution of technology, data, and science itself. While data centers remain a critical engine for cutting-edge research, there is an opportunity for them to do more in today's data-intensive, trans-disciplinary scientific world.

This paper describes a vision for "Data Center 3.0"—a new model for making data stores discoverable, accessible, useful, and applicable across all scientific domains. By identifying and tackling the fundamental challenges faced by today's data centers and their users, this model dissolves data silos and unleashes the power of data for future scientific breakthroughs and innovations.

The Challenge

Today's data centers that store and maintain large collections of scientific data are the products of a decades-long history at the confluence of science and technology. The model we call **Data Center 1.0** refers to the first generation of scientific data centers that emerged in the 1960s-1980s. While many of these first generation data centers are still in operation and have been upgraded to meet changing needs, when they were built they functioned essentially as archives for long-term data storage. Most are centralized in one physical location, managed by one organization, and designed for one research domain (see, for example, ICPSR, the Inter-university Consortium for Political and Social Research, established in the 1960s, http://www.icpsr.umich.edu/, and EROS, the Earth Resources Observation and Science Center, established in the 1970s, http://eros.usgs.gov).

The explosion of digital scientific data that began in the 1990s required a new approach, **Data Center 2.0**. Larger data sets, more complex data, and greater variety of data required scientists to think more strategically

An old school IBM data center from the era when computers used tapes.

AT A GLANCE

- Transformations in science and technology have brought new research opportunities and challenges.
- Scientific data centers must evolve to support more complex, innovative, and team-based science.
- A new data center model—Data Center 3.0—will enable researchers across scientific domains to turn raw data into real solutions for society's most pressing problems.

about how to manage and maintain the data, giving rise to new data management tools, metadata standards, interoperability protocols, and digital archiving standards. These more rigorous and complimentary standards in turn made it possible for the second generation of data centers to develop limited suites of tools to help researchers manipulate, analyze, and collaborate around data. Once again, centers established during the 2.0 period have continued to evolve as needs have changed and still function as important scientific resources. An example of a 2.0 center is the National Institutes of Health's National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov) that allows researchers in molecular biology and genetics to submit, discover, download, and analyze data in a single web-based framework.

Although both the Data Center 1.0 and 2.0 models have enabled substantial research advances, scientists are beginning to realize that they need new capabilities in the 21st century. The rise of big data and significant increases in computational power have been a key focus for data centers, but the data management and analysis tools needed to harness big data have been implemented inconsistently and with varying levels of success. Crucially, despite a drive toward multidisciplinary science, the vast majority of existing data centers still constrain data to domain-specific silos and retain a core model of physically centralized storage, access, and management. And despite policy and sociological trends toward open data, most data centers are still designed around a closed data lifecycle model in which data is generated and then submitted to an archive for formal curation. Data is generally held by a researcher or data creator, handed off to a data center, and only made available after ingest into the archive. The trend toward multidisciplinary, data-driven science and the need to exploit big data create a need, and an opportunity, to envision the data center's next evolution.

Technology, Data, and Science Are Changing

The practice of science has undergone remarkable changes since the advent of the data center. Classical science of hypothesis-driven, single investigator research morphed in the last half of the 20th century to what some have dubbed "Science 2.0." Science 2.0 is characterized by collaborative, multi-investigator, open integrative research (Shneiderman, 2008; Waldrop, 2008). With the advent of the data deluge, science is described as entering a Fourth Paradigm, data-intensive science. The practice of research has passed through phases of

direct observation, theoretical exploration, and data computation into an era in which science is about *data exploration* (Hey, 2009). Some view this shift as a scientific revolution akin to that described by Thomas Kuhn in his book *The Structure of Scientific Revolutions*, in which science is inductive, agile, and more outcome oriented as opposed to process-oriented. Unfortunately, science institutions, such as data centers, are often slow to react to emerging new realities.

In large part, these trends have been driven by advances in the technological tools that underlie scientific research, including sensor technology, data storage capacity, and computational horsepower. Increasingly sophisticated instrumentation enables scientists to collect data at unprecedented rates, often in real time. For example, the Sloan Digital Sky Survey (SDSS, http://www.sdss.org) telescope collected more data in its first few weeks of operation than had been collected in the entire field of astronomy up to that point (NSF, 2012). Sensor technology in satellites and ground-based stations allow the constant monitoring and measurement of the entire Earth. Personal devices, from life-saving (pacemakers) to lifestyle (smart watches), constantly collect data of potential value to research, medicine, business, and society.

> “Science urgently needs solutions that enable researchers to discover and share data and software, along with structures or incentives for actually repeating and reproducing studies.”

Distributed computation is often a convenient and cost-effective solution for individual researchers or labs to access greater computational power and storage capacity. But a lack of standardized architectures and procedures means that these systems are not currently being used as efficiently as possible. As a result, researchers are storing multiple copies of big data sets in various clouds and on local servers, thus adding to the amount and complexity of data that must be managed.

The increasing amount and complexity of data also creates a need for more sophisticated analytical tools and computational horsepower for manipulating and using data. Simulation and modeling, model validation, and model intercomparisons are now crucial aspects of the scientific endeavor. These processes require flexible access to varying amounts of computational power, highly sophisticated software, and specialized personnel to perform data modeling, develop and maintain software, and create data visualizations. No longer merely a means to an end, data and the software needed to process it are increasingly recognized as scientific contributions in and of themselves. Yet there are no established standards for documenting, evaluating, and archiving research software, and domain scientists are not always adequately trained to develop and effectively use the software they need.

In recent years, there has been a renewed push to reclaim the core elements of the scientific method through greater transparency, accountability, and reproducibility in research (Carey, 2015). A string of highly visible incidents involving scientific misconduct or outright fabrication have put in stark relief the shortcomings of the traditional peer-review system—a system that does not, in general, require researchers to provide their raw data or custom computational software, and in which few studies are ever reproduced. A lack of standard methods for archiving and sharing raw data, a lack of incentives for scientists to reproduce the work of others, and an explosion in the amount of data and the sheer computational muscle required to complete many analyses today have made it effectively impossible for the research community to routinely reproduce experiments and

analyses. Science urgently needs solutions that enable researchers to discover and share data and software, along with structures or incentives for actually repeating and reproducing studies.

Researchers today are awakening to a new way of thinking in which *everything* is a multidisciplinary problem. But while multidisciplinary work is now seen as crucial to solving society's critical challenges, carrying out truly multidisciplinary work is challenged by the fact that different disciplines collect, analyze, and assign value to data in different ways. When one domain sponsors or owns the data, that domain's research methods and data metrics tend to dominate the scientific findings (Stirling, 2014). The data owners remain the "data brokers" for their particular research community, keeping data isolated and inhibiting true interdisciplinary collaboration. To realize the full potential of the Fourth Paradigm, scientists need open data, better tools, and a cadre of data professionals dedicated to cleaning, describing, and accessing data to enable multidisciplinary work.

As scientific, policy, and sociological trends favor increased data sharing, it is also important to remember that data represents a valuable asset and that producing it can require tremendous time and energy. Sharing data raises concerns, for example, that researchers not familiar with the data will make false assumptions and generate flawed conclusions, or that "research parasites" may co-opt the research productivity of others for their own gain (Longo, 2015). Multiple models exist for providing access to data: sharing data freely, embargoing it for some period of time, charging for use, swapping data for other assets (including other data), requiring collaboration for use, and placing requirements around authorship. No one model can meet the needs of all communities or investigators, and new models may yet emerge. The next wave of data centers must balance the needs of both data generators and data users to establish a just and mutually-beneficial system for data exchange.

Critical Data Challenges in Science Today

1. ***Use of data across scientific domains:*** Researchers need tools to access *and* understand and appropriately interpret data collected by others.
2. ***Data sharing, with appropriate incentives and rewards:*** Researchers need to seamlessly access and combine multiple data sets. This requires consistent and recognized incentives to encourage data sharing and appropriate rewards for data generators.
3. ***Data management:*** Researchers need a new, more consistently implementable set of global metadata standards and interoperability protocols.
4. ***Code management:*** Software and code are increasingly integral to the scientific process and are increasingly integrated with the development of scientific data. Scientists need access to tools and methods to support sustainable scientific software.
5. ***Access to compute and data analysis:*** An increasing reliance on modeling, analytics, access to reliable analytical resources, and visualization requires access to compute power beyond the typical desktop. Researchers will increasingly rely on high performance computing infrastructure.
6. ***Publication and dissemination:*** Researchers need mechanisms for greater transparency, accountability, and reproducibility of research results.
7. ***Privacy:*** Sharing data and integrating data from multiple sources increases risks related to protecting confidentiality and privacy. In a more highly interconnected world of Data Center 3.0, the infrastructures must be created to maintain privacy.

8. ***Flexible networking architecture:*** *T*he integration of data from various sources and the need to bring data to compute or vice versa will place added demands on networking infrastructure. New approaches are being developed, such as software defined networking, that allow for more flexibility for creating on-demand networking infrastructure.

Data Center 3.0: A Call to Action

Data scientists and domain researchers at the Renaissance Computing Institute (RENCI) are establishing a vision for the next generation of scientific data centers called "Data Center 3.0." This vision is based on a series of principles that are crucial for data centers to continue to meet the needs of the research community across all scientific domains and across the entire data-collection-to-knowledge-generation lifecycle. By uniting existing and new community-driven, expert-mediated, multidisciplinary scientific resources into one integrated data network, we can build a more sustainable and useful data resource to fuel scientific advances.

Core Principles

1. ***Data Center 3.0 connects distributed data stores.***
Traditionally data centers comprised physically centralized computer systems with narrowly defined input and output channels that required users to download the data to their local computers and use their own computing power (and their own software) to work with it. In Data Center 3.0, we see each physical data center as a node within a larger "data grid" that users can tap into from anywhere. This grid will enable a more dynamic, multi-way exchange in which researchers will be able to rapidly access, transfer, and compute on large data sets from any location, making it easier for researchers in any field to use and share data.

Developing this data grid will require reliable middleware to organize and maintain a cohesive overarching structure, as well as robust networking capability to connect and handle massive data streams. Because the data collection nodes already exist as 1.0 and 2.0 Data Centers, these nodes do not need to be built and the storage medium and approach of each physical data center is irrelevant. The true innovation of the Data Center 3.0 is the grid that connects these data storage nodes.

Because scientists often need to access large volumes of data or integrate data from various sources on the fly, it will be important in some cases to containerize compute processes as virtual environments. Moving the computation to the data (or by moving the data to the computation) will allow much greater flexibility for the researcher and create the types of digital objects (data and code together) which will further support scientific reproducibility. Connecting these resources cannot be done in a static way and must rely on web services interfaces and other types of advanced networking protocols such as software defined networks. These elements will be a central part of a Data Center 3.0 implementation.

The RENCI data center today. RENCI is working to implement Data Center 3.0 strategies.

2. ***Data Center 3.0 has strong metadata standards.***

Metadata has long been a clear priority for data centers, and its thoughtful application has been enormously beneficial to all fields of science. However, there has been limited success in requiring individual researchers to do their own metadata input (Ramachandran & Khalsa, 2015). In addition, because most data centers have historically been domain specific, it is very difficult to develop a one size fits all approach to metadata. The increasingly interdisciplinary nature of science makes the use of metadata—and the development of common metadata standards—even more important even as the rapid increase in data volumes and complexities makes the generation of appropriate metadata even more challenging and critical.

Data Center 3.0 will use best practice metadata solutions and accepted structured metadata protocols to make data discoverable and accessible to users from diverse fields with diverse needs. These protocols will provide clarity and ease of use while reducing redundancy and confusion. Recognizing that creating accurate, useful metadata can require significant time and effort, Data Center 3.0 should provide tools to automate the creation of metadata to the extent feasible. By providing automatic metadata suggestions that researchers can then accept or modify, by encouraging the integration of metadata creation into the research process, and by providing meaningful incentives for data generators to properly annotate their data, Data Center 3.0 will support developing the required metadata while lowering the burden on data contributors.

3. ***Data Center 3.0 data is (more) interoperable.***

With the exception of large data campaigns such as Earth observing satellites, most data has been generated and housed organically, in local ecosystems and according to local practices that often vary from researcher to researcher and discipline to discipline. This contributes to walled gardens of data and limited means to promote interoperability.

To advance beyond the current state in which each scientific domain has its own local vocabularies, tools, themes, software, and applications, Data Center 3.0 will link data and support cross-disciplinary data access and analysis. Although it may be impossible to achieve true interoperability across federated data collections, Data Center 3.0 distributed data grids, metadata protocols and semantic interoperability, and linked data can improve the range of options and help to create a more workable path forward.

4. ***Data Center 3.0 includes robust analytical tools.***

Today's increasingly interdisciplinary, data-intensive, software-reliant scientific landscape requires new analytical tools to make sense of the vast streams of data we can now generate. Many of these tools already exist, but are isolated within data centers or scientific disciplines, accessible only to a tiny fraction of researchers who could potentially use them. Data Center 3.0 will make it easier for researchers to find and use data processing tools in addition to data itself. By using analytical tools through a data grid, scientists will be able to analyze multiple data sets, perform real-time computations, and create simulations without the constraints of their own computational power or local hardware storage space.

To support the creation and use of these data analysis tools, Data Center 3.0 will be built with layers of adjustable, reconfigurable network services. A suite of standardized, configurable tools will be designed to work broadly across all the fields of data, but may also be customized as the research demands. "Analytics-as-a-service" is a useful model for the kind of middleware construction that can be built, offered, and personalized as needed. Data Center 3.0 will provide appropriate recognition and rewards for tool contributors.

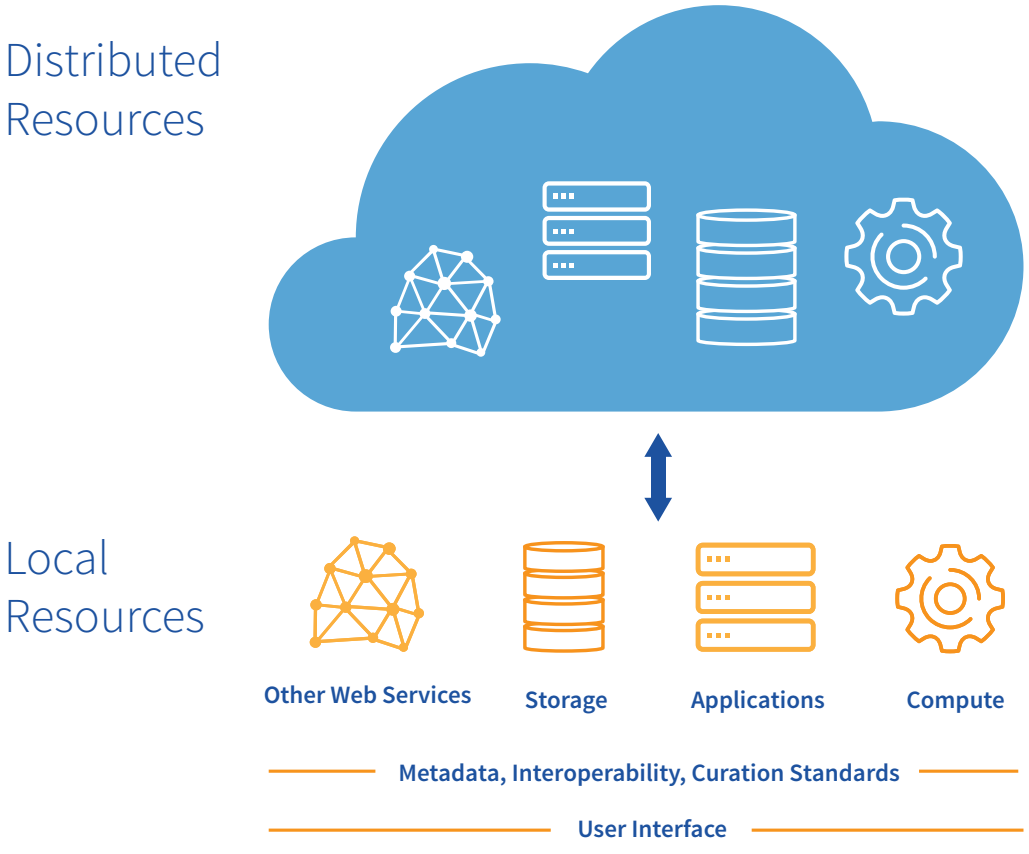

FIGURE 1.
Data Center 3.0 provides seamless integration of local and distributed data and cyberinfrastructure for end users.

Existing Efforts that Exemplify Data Center 3.0 Principles

Data Center 3.0 is not a completely new idea, but rather an extension and unification of principles already percolating in the data science community. It is useful to explore how these principles are being applied or attempted in existing efforts as we look toward a more universal implementation across the entire research world.

Community-driven

Community is essential in scientific work today, and is becoming a cornerstone of new data infrastructure. The EarthCube Project (http://earthcube.org) is an excellent example of a community-driven data resource. Membership is free, and users are able to access computing power, multiple data sets, and powerful tools to share, analyze, and visualize data in the geosciences. Another robust data sharing community is CyVerse (formerly iPlant, http://www.cyverse.org). CyVerse users can access a powerful computational infrastructure that enables sharing, analysis, and collaboration for huge, complex data sets in the biological sciences.

Expert-mediated

The trained experts that manage data centers in the 1.0 and 2.0 models provide crucial services for researchers. NASA data archives (https://earthdata.nasa.gov), for example, are well-known for such experts, and each division has standards in place in order to anticipate, understand, and meet their users' needs (Mehrotra et al., 2014). In addition, NASA has long appreciated that data must work together and has long been part of a confederation of world data centers. The Data Center 3.0 model can build on and expand these successful models to encompass all scientific domains.

Multidisciplinary

There have been several initiatives to connect individual data sets into larger data resources that span multiple disciplines. For example, the formerly-separate National Climatic Data Center, National Geophysical Data Center, and National Oceanographic Data Center were transformed into the National Center for Environmental Information (https://www.ncei.noaa.gov) in part as it became clear that researchers from across the earth sciences needed better tools to combine and compare data from across these related fields. Similar efforts in different disciplines include CyVerse, iMicrobe, and the planned Virtual Institute for Social Research.

We believe Data Center 3.0 has the opportunity to build on these successes and address the needs of the scientific community and society as a whole.

5. ***Data Center 3.0 is agile, user-driven, and expert-mediated.***

The typical top-down, centralized data management structure of the Data Center 1.0 and 2.0 models provides numerous benefits, such as relatively stable funding, persistent and consistent data collection, enforcement of local metadata standards, and the availability of domain expertise. That centralized approach, however, has also led to the establishment and persistence of data silos and creates a system in which change is expensive and slow.

	Attribute	Data Center 1.0 (past)	Data Center 2.0 (current)	Data Center 3.0 (future)
Physical Cyberinfrastructure	*Infrastructure*	Centralized	Centralized	Distributed
	Storage Capacity	Limited, dependent on data center resources, physical media	Centralized by data center, dependent on data center resources	Distributed, potentially unlimited in terms of capacity
	Computation	Limited to user's own resources, done on user's own system	Limited number of shared tools available, (e.g., webGIS), mostly done on user's own system	Agility in using local resources and community-developed, cloud-based services available in the data grid
	Networking	Bitnet, Internet	Internet2	Software Defined Networking
User Interfaces	Discovery	Domain knowledge, hard copy catalogs, beginnings of basic online catalog systems	Websites run by each repository; sophisticated digital data catalogs	Custom search algorithms and recommendation engines
	Access	User accesses data on physical media, obtains "complete" dataset	Internet-based distribution; user must access each data collection separately from host data center	Federation of data systems
	Protocols	FTP, Globus, some web services, some middleware	Open APIs, web services, advanced middleware	Seamless interoperability, curation, and compute
Interoperability, Standards, and Curation	Metadata Standards	Nascent standards	Determined by domains and research communities	Standardized and automated via middleware
	Data Curation	Driven by data "owner"	Open Archival Information System (OAIS), ISO 16363	Agile, community-based; mix of decentralized and centralized, curation throughout data life story
	Software Curation	N/A	Beginnings of efforts to archive scientific models	Open code development, community-based sustainability; increasing linkages between data and software
	Publication - Dissemination	Centralized at data center	Centralized at data center	Integrated into data grid; involves data centers, journals and other relevant institutions

TABLE 1.
Key attributes of Data Center 3.0 compared to past stages in data center evolution

We envision Data Center 3.0 as an agile, user-driven endeavor that provides appropriate incentives and rewards for users from all fields to input, work with, and collaborate around data. Being agile will allow the data network to evolve quickly to stay ahead of the demands of its users. Similarly, being user-driven makes it more likely that Data Center 3.0 will be user-friendly and truly meet the needs of the research community. The user's expectation is that data management tools be simple and seamless (Ramachandran & Khalsa, 2015). User authentication and access must be swift and easy.

At the same time, we believe Data Center 3.0 must be expert-mediated by a core staff, representing data science and curation experts with domain knowledge, who can be on hand to guide users through data discovery, facilitate the use of new analysis tools, personalize data services based on research preferences, and drive collaboration across domains.

6. ***Data Center 3.0 facilitates community-driven science.***
Data centers of the past created and supported a community of researchers who were engaged in one scientific realm. As we have seen, the scientific pursuit is far more integrative now. Research questions and potential solutions combine many fields, can benefit from many different data sets, and may have surprising applications in other domains. Data Center 3.0 will facilitate community-driven, multidisciplinary science through tools that unite its users, instead of funneling them into domain silos.

The Upshot

With Data Center 3.0, we can create a new umbrella infrastructure for scientists to capitalize on opportunities in data-intensive science. Implementing this vision will not require a wholesale replacement of the current data center model. Rather, Data Center 3.0 will build upon the excellent and time-tested ideas and structures that science and data centers have already established, while moving forward to address key challenges. This new infrastructure will facilitate true interdisciplinary work among researchers, data managers, and software developers and "set data free" in the process.

In addition to substantially greater utility, we believe the Data Center 3.0 model offers a *sustainable* path forward. Because the infrastructure will be widely distributed and cloud-based, Data Center 3.0 will allow for greater efficiency in data storage, more optimal allocation of computing resources, and lower overall costs and energy consumption. In addition, a shared system will use personnel more efficiently and allow different resources to be updated at different times, so that the system is never completely offline. In short, Data Center 3.0 will pave the way for better data, better software, and better science.

The Big Picture

Data for 21st-century science deserves new tools that can respond to the needs of the ever-changing technology, data, and science itself. With Data Center 3.0, researchers will be able to tap into a resource that is vastly larger, more diverse, and more powerful in order to meet the demands of data-intensive science.

Data Center 3.0 offers a bold vision to integrate scientific data across all its fields. The result will be more and better research for each research dollar invested, new scientific discoveries, new technology products, and new solutions for society's most vexing problems.

REFERENCES

Carey, B. (2015, June 15). Science, Now Under Scrutiny Itself. *The New York Times*. Retrieved June 28, 2015, from http://www.nytimes.com/2015/06/16/science/retractions-coming-out-from-under-science-rug.html?_r=0

EarthCube. (n.d.). Retrieved September 19, 2015, from http://earthcube.org/home

Hey, A., Tansley, S., & Tolle, K. (Eds.). (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, Wash.: Microsoft Research.

iPlant Collaborative. (n.d.). Retrieved September 19, 2015 from http://www.iplantcollaborative.org/

Mehrotra, P. et al. (2014). Supporting "Big Data" Analysis and Analytics at the NASA Advanced Supercomputing (NAS) Facility. *NAS Technical Report, NAS-2014-02*. Moffett Field, CA: NASA Ames Research Center.

National Centers for Environmental Information (NCEI). (n.d.). Retrieved September 19, 2015, from https://www.ncei.noaa.gov

National Climate Data Center (NCDC). (n.d.) Retrieved September 19, 2015, from https://www.ncdc.noaa.gov/about

National Science Foundation. (2012). Cyberinfrastructure Framework for 21st Century Science and Engineering. Retrieved June 27, 2015, from http://www.nsf.gov/cise/aci/cif21/CIF21Vision2012current.pdf

Ramachandran, R., & Khalsa, S. (2015). Moving from Data to Knowledge: Challenges and Opportunities [Technical Committees]. *IEEE Geoscience and Remote Sensing Magazine, 3*(2), 51-54.

Shneiderman, B. (2008). Science 2.0. *Science, 319*(5868), 1349-1350. Retrieved from http://www.jstor.org/stable/20053525

Stirling, A. (2014, June 11). Disciplinary dilemma: Working across research silos is harder than it looks. *The Guardian*. Retrieved September 19, 2015, from http://www.theguardian.com/science/political-science/2014/jun/11/science-policy-research-silos-interdisciplinarity

Waldrop, M. M. (2008). Science 2.0. *Scientific American, 298*(5), 68-73. Scientific American, Inc. Retrieved from http://dx.doi.org/10.1038/scientificamerican0508-68

ABOUT RENCI:

RENCI, an institute of the University of North Carolina at Chapel Hill, develops and deploys advanced technologies to enable research, innovation, and economic development. For more information, see www.renci.org.

HOW TO REFERENCE THIS PAPER:

Lenhardt, W. C., Blanton, B., and Schmitt, C. (2016): From Data Center 1.0 to Data Center 3.0: Transforming the Storage, Access, and (Re)Use of Research Data for Better Science. RENCI, University of North Carolina at Chapel Hill. Text. http://doi.org/10.7921/G04M92GJ

renci

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL