Name It! Store It! Protect It!: A Systems Approach to Managing Data in Research Core Facilities

Matthew DeVries,¹ Matthew Fenchel,² R. E. Fogarty,¹ Byong-Do Kim,³ Daniel Timmons,⁴ and A. Nicole White²

¹Case Western Reserve University School of Medicine, Cleveland, Ohio 44106, USA; ²Cincinnati Children's Hospital Research Foundation, Cincinnati, Ohio 45229, USA; ³University of Virginia, Henrico, Virginia 23294, USA; ⁴Biofrontiers Institute, University of Colorado, Boulder, Colorado 80309, USA

As the capabilities of technology increase, so do the production of data and the need for data management. The need for data storage at many academic institutions is increasing exponentially. Technology is expanding rapidly, and institutions are recognizing the need to incorporate data management that can be available for future data sharing as a critical component of institutional services. The establishment of a process to manage the surge in data storage is complex and often hindered by not having a plan. Simple file naming—nomenclature—is also becoming ever more important to leave an established understanding of the contents in a file. This is especially the case as research experiences turnover from research projects and personnel. The indexing of files consistently also helps to identify past work. Finally, the protection of the data contents is becoming increasingly challenging. As the genomic field expands, and medicine becomes more personalized, the identification of methods to protect the contents of data in both short- and long-term storage needs to be established so as not to risk the potential of revealing identifiable information. This is often something we do not consider in a nonclinical research environment. The need for establishing basic guidelines for institutions is critical, as individual research laboratories are unable to handle the scope of data storage required for their own research. In addition to the immediate needs for establishing guidelines on data storage and file naming and how to protect information, the recognition of the need for specialized support for data management supporting research cores and laboratories at academic institutions is becoming a critical component of institutional services. Here, we outline some case studies and methods that you may be able to adopt at your own institution

KEY WORDS: data management, data sharing, data security, file naming

NAME IT!

The discussion of file naming probably sounds about as dry, technical, and—dare we say—“boring” as topics come. However, the development of a system of file naming can actually save much frustration and tedious searching at critical times in our work. It also makes the naming of a certain file much easier, as there is a predefined algorithm to follow.

The end goal of good file naming is simple: to be able to recognize at a glance the content of that file.

Take, for example, the words you have already read in this article. Without having to think about what each letter in a word meant, nor sound out the word itself, you knew at a glance what each word represented. Or, if you look at most photos, you can easily recognize what is in a picture without

needing to study every part of it. The goal of a file-naming system is the same: one look at the file name, and you basically know the content/metadata.

What is outlined here is one type of system. But, it is certainly not the only system that can be developed. Hopefully, you can glean the principles of a good file-naming algorithm and develop something that works intuitively and well in your work.

Some of the benefits of good file naming include, but are not limited to, the following:

- avoidance of unnecessary duplication of files;
- obvious file content to you and others;
- elimination of file loss;
- more efficient sorting and searching;
- avoidance of accidental overwriting;
- obvious file location.

As you are developing your own system, the following tips may be useful.

1. Develop standard terms, spelling, and capitalization that you will consistently use. Keep them short, and abbreviate where possible. For example, in the statistical

*ADDRESS CORRESPONDENCE TO: Nicole White, ML4010, 3333 Burnet Avenue, Cincinnati, OH 45229, USA. (E-mail: amanda.white@cchmc.org)

doi: 10.7171/jbt.17-2804-003

world, everyone knows that “SAP” stands for “statistical analysis plan,” so we use that abbreviation in our file naming.

2. Don’t try to duplicate all folder information within file names.

3. Don’t try to put into file names what should be in the file header. You don’t need all metadata about a file in the name.

4. Be systematic and consistent on use of special characters and spaces.

5. Terms in the name should follow the following:

- a. More important → less important, and/or
- b. General → specific, and/or
- c. Constant → changing.

We generally try to use four different levels of information in our file-naming scheme to meet point 5 above.

- LEVEL 1 – Largest unit of information (not contained in a folder name).

—Examples: Lab name. Study/Project name. Development or Production.

- LEVEL 2 – Content type.

—Specifications (Specs). SAP. Output. Report. Program. Data.

- LEVEL 3 – Short description.

—“H.1 Poisson.” “Derived.” “Original.” “Descriptive Statistics.” “CRF 1.”

- LEVEL 4 – Version.

—20160422. V01. SEPT2013.

—Recommend using terms that are sequential in nature. The use of the word “final” in a file name can be problematic. (What if there is really another version after the final one?)

At this point, we thought we would share some “poor” examples of file naming, along with the “new and improved” version below in italics. To “protect the guilty,” we have altered these examples for the purposes of this article.

- analysisoutput.doc

—*PFT Output H.1.2 ANOVA 20161122.doc*

- Mar09 magnetic resonance imaging comparisons from Johnson lab.csv

—*JohnsonLab_Report_MRI_Analyses_MAR2009.csv*

- Specs (derived variables) v#1.rtf

—*MI_Study_Specs_Der_Var_V01.rtf*

- program improvecarenow descriptive.sas

—*ICN Program H.1.1 Descr Stats 20140814.sas*

- Raw data in kg, pg/ml, by event date final.txt

—*IBD Study Data Orig V298.txt*

Example a did not provide enough details and was difficult to read without some type of spacing and capitalization. Example b did not follow a logical order,

could have used “MRI” as an abbreviation, and did not need the word “from.” Example c had special characters; although not wrong, they may not be able to be read by all computer systems. Example d, again, did not have a logical order or enough information. Example e listed units of measure, which normally are not necessary.

Finally, we try to think of file names in the sense of “titles,” such as a title on an article. Titles are typically not actual sentences, but phrases. Titles typically capitalize all initial letters in words (except words, such as “the,” “and,” *etc.*). Manuscript titles give enough key information—without being too long or too short—to know what the content will be about.

Final, yet vitally important considerations, are the limitations of the operating system (Windows, Linux, Sun, iOS, *etc.*), 1) where the files are being generated/stored and 2) where the files will eventually be used (if different from the original operating system). Depending on the operating systems involved, the following may need to be considered in naming conventions:

- length of name;
- use of spaces and/or the underline character;
- use of numerals;
- case (upper, lower);
- use of special characters (–, +, #, *etc.*);
- extensions (*.doc, *.txt, *.exe, *etc.*).

Planning files names in a way that the portability/ interoperability of the files between operating systems is in place can save much time and frustration.

Again, in the end, you want file names where, at a glance, a person will recognize the content of the file. Develop a system that works for you and your team using terms that you are familiar with that can be easily sorted and easily searched. Once you have a system in place, be deliberate about its consistent use. Although this may be tedious and mundane at the beginning, eventually, it will become habit, saving everyone time and effort in the long run.

STORE IT!

Many life science research organizations have looked into possible solutions in enterprise storage solutions with advanced data-protection features, such as fail-over configuration, clustered file system, and several backup methods. Although they provide multiple data-protection layers, as a result of their prohibitive cost, it would not be possible to create suitable storage solutions for ever-increasing biomedical research data (Fig. 1). In the science and engineering research community, high-performance computing (HPC) and storage systems have been used for large-scale computational simulations and big data management. The

FIGURE 1

Storage systems and data workflow at UVA.

practices in the HPC community can be adopted for data storage and management in biomedical research, and the recent advancement in high-performance parallel file systems and scale-out network-attached storage (NAS) storage systems makes it easier to design storage solutions specialized for life science and clinical data.

In this section, 2 case studies of storage systems implementation at the University of Virginia (UVA) and the University of Colorado at Boulder are presented. Both have the same goal of supporting research data storage and management using HPC resources.

At UVA, the School of Medicine Research Computing works with the central information technology (IT) organization to create an efficient data storage and management environment for researchers. **Figure 1** shows UVA's HPC, secure data analytics platform, and storage systems implementation on the campus. Here, we will focus on illustrating the storage system setup. The numbers in each section in the diagram indicate different stages of different storage systems, while research data are transferred depending on the needs of researchers. Below is a short description of each stage.

1. Nonsensitive data generated from research labs and cores can be directly transferred to either the HPC parallel file system (Lustre) or the long-term NAS system. Automatic transfer from lab instruments to the long-term NAS storage (2 petabytes) *via* Science DMZ data transfer node (DTN).

2. Once the data are processed on the HPC system, the postprocess data can be transferred to the long-term NAS storage. Users can have an automated daily backup cron job running in background or do the transfer manually before the 90 d data purging policy kicks in on the Lustre file system.

3. For users who need a backup solution for their data, another fast and cheap storage option is available connected to the long-term NAS storage. The object storage has been implemented as a semiarchiving system. The system is 100% disk, which guarantees faster operation than the

conventional tape-archiving system, and the technical design of object storage allows massive scalability.

4. UVA's central IT also offers an enterprise level of data protection with different arrays of storage systems. Users can pick and choose various combinations of different data protections, such as 2 wk snapshot, replication of data at a different location for disaster recovery, and a tape-archiving system.

5. Sensitive data from the patient data warehouse at the university hospital can be extracted for clinical research and stored in a separate storage system that satisfies the Health Insurance Portability and Accountability Act compliance in terms of security and data protection. The secure storage is mounted on the secure data analytics platform.

6. In addition to the current storage implementation on campus, UVA is planning to deploy another system that provides application programming interface programmability. This new storage will be the gateway storage system for data sharing with collaborators outside of UVA and will also connect to Cloud infrastructure for collaborative research projects among multiple institutions that require a Cloud platform with a more complicated design.

The BioFrontiers Institute at the University of Colorado has an environment custom tailored to the specific needs of a group of researchers with diverse specialties. These resources have been selected to complement already-existing campus infrastructure and provide enhancements where BioFrontiers investigators have demonstrated unique requirements, not always well served by standard campus systems. **Figure 2** shows a conceptual layout of research instruments, storage systems, and HPC infrastructure designed to meet these needs.

Our workflow includes the following:

1. Scientists collect data from a variety of core facilities that can include live-cell imaging data over extended time periods or large sequencing runs. These data streams easily contain several terabytes of data per experiment, and advances in imaging technology and automation are constantly expanding the size of datasets.

FIGURE 2

Storage systems and data workflow at the BioFrontiers Institute.

2. Users directly manipulate and curate research data on the lab storage from individual workstations and prepare data for movement to HPC systems.

3. User workstations and instruments connect to DTNs to transfer data to a variety of storage resources, internal and external to the BioFrontiers computing environment.

4. Raw data are moved to the parallel file system in the HPC environment, where analysis is performed. This step can be done directly from mounted client machines or using DTN systems when larger transfers are required.

5. For distribution and publication of primary research, data cluster storage is connected to a virtualized web cluster that can be used to develop platforms and tools to distribute data widely.

6. To use campus resources and share data with collaborators, data are moved from the BioFrontiers environment to campus resources with a more general user base.

7. As commercial cloud integration becomes more mainstream for research purposes, DTN systems are able to integrate directly with a variety of available solutions.

PROTECT IT

“Data security is a journey, not a destination,” proves true, particularly in today’s rich data environment for core research services. The size and scope of data sets continue to

grow exponentially, and the need to collect, store, share, and transfer data often exceeds expectations. In response, academic research institutions scramble to create environments to satisfy space and data analytics needs. Moreover, institutions face compliance requirements from state and federal entities that invoke information security best practices, as well as *via* contractual terms with external academic or nonacademic partners.

The implications of these requirements for human subject research are substantial. Additional focus on data protection and consultation with information security professionals are now absolute necessities. Furthermore, to ensure the human subject research protections outlined in the Belmont Report and Common Rule are met, the researcher, now, must consider a real and tangible risk to a consented subject’s confidentiality and/or electronic presence.

Academic research institutions, core directors, and researchers lean on their IT Departments and Information Security experts to propose or establish security controls and best practices for data-protection requirements listed in data use agreements with other institutions, Institutional Review Board (IRB) protocols, contracts, and grants.

Academic researchers and information security specialists have always collaborated on security controls and best practices for IRB protocols. In this new environment, with the increase in data complexity and regulatory oversight, the seasoned researcher and IT specialist may find that necessary

protections impact methods of research to the extent of potentially altering original research questions or methods for a safer or less-expensive collection or storage method.

For example, use of an iPad to collect identifiers, such as name, date of birth, diagnosis, *etc.*, may now require mobile-device management controls, potentially increasing the cost to the researcher.

The core director and/or principal investigator (PI) may edit or change their research method to match security controls, which may impact the value of their original research question/methods in favor of a safer or less-expensive collection or storage method.

Therefore, the governance model designed to enhance data security is now being adapted to research. The details that a core director or PI needs to determine when designing a human subject research protocol involve identification of data ownership; size of data at the start *vs.* after big dataset transfers; controlled list of authorized users; and evaluation of the environment for collection, transfer, sharing, analysis, storage, and destruction. In response, IT Departments work arduously to create safe environments for the data. For instance, Case Western Reserve University's Institute for Computational Biology and UTech (IT) Department created a Secure Research Environment as a Health Insurance Portability and Accountability Act- and Federal Information Security Management Act-compliant location for research data. The cores can now provide storage locations that may qualify as a HPC center and/or a secure research environment for electronic medical records or large images that include identifiable information.

IT Departments assist the cores and/or PIs to triage campus data *via* data-sensitivity categorization and assign datasets and enterprise software to secure, semi-secure, and open environments. This adjustment needs to be handled with sensitivity, as higher education research institutions are, by intent and design, open environments where sharing and collaboration are the expectation.

When handling regulated data, Case Western Reserve University's IT specialists call upon the National Institute of Standards and Technology special publications 800-53 and 800-171, IRB confidentiality protections (Belmont Report and Common Rule), and requirements for a genome-wide association study or genotype information to formulate a standard approach to protecting data.

DISCUSSION

The implementation of a cohesive data management plan at your institution is better served as a proactive attempt *vs.* a reactive scramble. There are three major components to ensuring that research data from core facilities and laboratories are cohesively managed and guaranteed for future use of downstream research. Each of these

components—data storage, file naming conventions, and data security—addressed is better served and managed as a centralized resource in an institution. The management of these data as a centralized, institutional resource will allow for a cohesive process to be established, the implementation of policy to be communicated and executed more effectively, and the identification of data management solutions and troubleshooting to be conducted more swiftly.

This paper is established as a suggestive framework built on case studies from institutions that are actively adopting data management workflows and solutions. Research agencies are requiring data management plans and the release of data sharing under some terms of funding.

Many of the funding agencies are requiring a description in the data plan of the file types, a data storage plan, how data will be shared with others, how long-term preservation of files will be maintained, and when data will be made available to the public. A composite of a few of these U.S. agencies requiring this information is listed below:

- NIH https://grants.nih.gov/grants/policy/data_sharing/
- NIH Genomic Data Sharing <https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/>
- National Science Foundation <https://www.nsf.gov/bfa/dias/policy/dmp.jsp>
- Department of Energy <https://science.energy.gov/funding-opportunities/digital-data-management>
- National Aeronautic and Space <https://www.nasa.gov/open/researchaccess/data-mgmt>
- Department of Defense http://www.dtic.mil/dtic/pdf/dod_public_access_plan_feb2015.pdf

More resources can be found across the web as it relates to the type of research a core facility or laboratory may be focused on. Standards are being established across the scientific disciplines as well. Although this paper does not go into the expectations of digital information contained within the files, it is something that each core facility should take into account when creating data for researchers within its institution to ensure researchers comply with publishing and funding agency standards.

Each of these authors would agree that the creation of a data management plan for an institution is not an easy feat. However, the ability to create a centralized process and to ensure that the management of research data is being addressed is a key and critical need for institutional oversight and to ensure that research data are protected and available for future use.

ACKNOWLEDGMENTS

The authors would like to acknowledge Matt Hynes-Grace at the Biofrontiers Institute and University of Colorado for contribution to the figures included in this paper and editing its content. We would also like to acknowledge Heather Richards at the Comprehensive Cancer Center, The Ohio State University on her contribution to this work.