James L. Harroun. Investigating a Research Environment for Structural Bioinformatics Research Data. A Master's Paper for the M.S. in I.S. degree. November, 2013. 57 pages. Advisor: Stephanie Haas

This study investigates the Structural Biology Central Facility at a large academic research institution and examines the Central Facility's most pressing data management needs. The University's Health Affairs Library has undertaken an initiative to pursue data management support for researchers in the biomedical sciences and the central facility has been identified as one of the library's first partners in assessing and proposing data management services the library could provide.

Five groups of campus stakeholders that mutually contribute and participate in assisting the structural biology central facility were identified and interviewed to identify contributions to the central facility's data management practices. From these interviews, seven major obstacles were identified around which the Health Affairs Library could develop novel strategies for extending data management services into biomedical research environments.

Headings:

Science -- information services

Information science

Library extension

Bioinformatics

Big data

INVESTIGATING A DATA MANAGEMENT ENVIRONMENT FOR STRUCTURAL BIOINFORMATICS RESEARCH DATA

by
James L. Harroun

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

November 2013

Approved by

Stephanie Haas

Introduction and Background

Digital information resources and networks of electronically accessible publications continue to evolve into richer, more timely, and more easily accessible systems that disseminate knowledge through ever-evolving infrastructures (technical and otherwise), within emergent communities of use, and across diverse user populations. Dynamic and effective portals of access afford users efficient and timely access to publications and to data resources that are simultaneously dependent upon the state of the art of digital information networks and, through implementation and use, determinant of the digital networks through which they are provided. The visibility of digital information networks that disseminate knowledge is especially apparent in academic institutions and, in particular, academic institutions that emphasize research, scholarly publication, and that seek to identify challenging and innovative research questions and the breakthrough methodologies for answering them. Although networks of digital information resources afford more efficient and productive exposure to diverse information sources, the dependence upon them—and the ever-changing nature of this dependence upon them—continually poses new questions and challenges for information professionals, librarians, and libraries and their roles in providing expertise and support (Hey & Trefethen, 2003; Hey & Hey, 2006; ARL, 2007; Gold, 2007; Lewis, 2010; Corrall, 2012).

The reliance upon digital information networks and electronic publications continues to bring the roles of traditional libraries as places of information-seeking activities into question; these digital resources also continue to question and change the responsibilities of information professionals and librarians to best serve their current users and to develop innovative new services in support of emergent communities as the communities' complex information needs materialize. E-Science, sometimes referred termed cyberinfrastructure, has emerged as a burgeoning concept in which wet-lab and hands-on discovery practices are augmented and completed by in-silico experimental practices that project likely outcomes and confirm the validity of original experimental findings (Lyon,2008; Lyon, 2009; Lyon, 2012).. E-Science is founded on systems and structures of manifold data-processing devices, specialized computational applications, robust server and intercommunication networks, and electronic data stores (Newman, 2003; Jankowski, 2007; Stewart, Simms, Plale, Link, Hancock, & Fox, 2010). E-Science and research data management are emerging as vital opportunities for libraries, librarians, and information professionals. The challenges of e-Science emerge at the intersection of the complex and dynamic benefits and challenges presented by the increasing dependence on digital information, by the growing reliance on digital publication, and by the ever-increasing size and demands of the backbone of these publications: research data (Jankowski, 2007; Pace, Bardzell, & Fox, 2010; Bietz & Lee, 2012).

E-Science poses a rich and intense combination of computational, infrastructural, and domain-knowledge-dependent demands that extend beyond the academic library's practices: managing E-Science will most likely require scientific interdisciplinary collaboration, innovative technological partnerships, and the integration of the disciplines of Information Science and Library Science (ARL, 2007; Hey & Trefethen, 2003, Hey & Hey, 2007; Gold, 2007; Lyon, 2012). The disciplines of Information Science and

Library Science provide unique perspectives that extend beyond specific disciplines and practical contexts: academic libraries, librarians, and information professionals possess invaluable knowledge and skills to aid in discerning the lineaments of E-Science, to evaluate existing data management conventions, and to propose effective methods and practices to help mitigate many of the information demands that E-Science will generate(ARL, 2007; Hey & Trefethen, 2003, Hey & Hey, 2007; Gold, 2007; Lyon, 2012).

Data and, in particular, large sets of data, pose one of the most significant challenges for E-Science and effective research data management. Current Structural Biology and other computationally-based life science research is particularly data intensive, often demanding tremendous computational and data storage resources that are often substantial obstacles in effectively processing, storing, moving, and sharing data (Karasti, Baker, & Halkoa, 2006; Nam, Lee, Hwang, Suh, & Kim, 2008; Lyon, 2010). These data obstacles exist not only during the investigation and research phases. The immense amount of data generated from research projects can also pose longitudinal and, often unanticipated, difficulties in effectively managing and archiving research data after research has been concluded (Hedges, 2007; Lyon, 2009; Lyon, 2010; Pace et al., 2010; Kowalczyk, 2011; Stewart et al., 2010, Lyon, 2012) . In many ways, it is difficult to anticipate the current requirements and the long-term demands caused by the management of such large amounts of data. While this, at first glance, might appear to be a challenge that larger computational resources could solve, the issues at hand are much more complex and there are many nuances of managing these data that are difficult to anticipate and that continue to emerge as research practices and conventions evolve

(Kowalczyk, 2011; Bietz, Ferro, & lee, 2012). In many ways, the crises of large-scale data management mirrors the ongoing challenges that academic libraries, in particular, continue to face as digital information resources, and the reliance upon them, continues to evolve (Barga, Fay, Guo, Newhouse, Simhan, Szalay, 2008; Barateiro, Borbinha, Antunes, & Freitas, 2009) .

This study investigates a Central Facility focused on computational research in Structural Biology at a research-oriented academic institution. The study was initiated by the university’s Health Affairs Library as an investigation into how the Library can assist in the establishment of proof-of-concept practices that can assist in developing effective E-Science and research data management practices. The Structural Biology Central Facility is unique in the computational services it provides, its computational and data storage needs, and its position within this university’s organizational structure. A group of twenty-one subjects were interviewed with regard to their relationships and interactions with the Central Facility, to their research practices, to their data management practices, and to their understandings of E-Science, data archiving/sharing, and collaboration. The interviewees represent a broad population of principal investigators, graduate students, post-doctorates, administrators, and technicians who, when viewed as a whole, represents the Structural Biology Central Facility’s position within the university research community and captures the breadth of the central facility’s current practices, its interactions with various stakeholders, its current challenges, and its potential future challenges.

While many of the challenges faced by the Structural Biology Central Facility center around technological infrastructure, computational functionality, and challenges caused by the complexities of the research, a significant component of the Central Facility's data management challenges are based upon much different obstacles, such as organization of information, effective device and storage management, consistency of workflows, efficient system interactions, and communication among stakeholders. These challenges provide an excellent opportunity for the Health Affairs Library to leverage its strengths and current practices to address many imminent e-Science and data management challenges. Furthermore, information gained from this study will be used to assist the Health Affairs Library to better understand its current position and future plans and how the Library can act as an agent to identify potential campus collaborators.

Method

Five groups of stakeholders were identified as central to this study and as the groups of stakeholders who have the most frequent interaction with the Structural Biology Central Facility: the director of the Structural Biology Central Facility, the faculty principal investigators in the biomedical domains for which the Central Facility provides most consistent ongoing assistance, the post-doctorate and graduate students who perform the data processing and analysis under the principal investigators, the members of the university Research Computing division that provides and maintains computational resources, server access, troubleshooting, and initial training for those utilizing the computational research infrastructure, and the administrators and staff of the Health Affairs Library who are pro-actively seeking a means of involvement in the data worlds of those who utilize the Structural Biology Central Facility.[1] While the groups are ultimately involved in fulfilling successful e-Science endeavors, none of the groups, with the exception of the principal investigators and the post-doctorates and graduate students who work as a unit, are necessarily directly linked. That is, it is not necessarily the case that members of one principal investigator's lab will be aware of another principal investigator's research or data management practices, nor is it the case that members of the Research Computing team have ongoing relations with members of the health affairs library staff and administration, and so forth. In total, twenty-one subjects were chosen across the five groups identified. The director of the Structural Biology Central Facility was interviewed singly, seven faculty and principal investigators were interviewed, eight from the group of post-doctorates and graduate students were included in the study, two

[1] This study was approved by the UNC Chapel Hill Institutional Review Board

members of the research computing division, and three members of the Health Affairs Library staff and administration were included. Interview guides were prepared specifically for each group of interviewees and interviews were recorded, transcribed, and coded for qualitative analysis. Because this study is an initial investigation, the practical differences and variability identified among different groups of principal investigators, post-doctorates, and graduate students precluded quantitative analysis of the findings. As this is an initial investigation, the interview process spurred new questions and new issues arose as interviews were conducted.

Results

This study attempts to more finely locate the complex data issues at hand as well as offer a foothold for the Health Affairs Library into how the library's expertise can be provided as means of extending research data support into the biomedical sciences. Entering into the study, there were several assumptions about the set of participants investigated. These assumptions were based on a preliminary understanding of how the participants interact with one another and were, to a great deal, informed by existing literature on e-Science. Most assumptions focused on data size being a limiting and overwhelming factor for all involved and that the sheer volume of data—measured in disk space required to hold this data as well as the number of files to be maintained and organized—would be the Achilles heel of the central facility's data management issues as well as the most noticeable set of difficulties for the principal investigators, post-doctorates, and graduate students. Another assumption entering into the study was that access restrictions and data rights would be identified as major issues that would need to be addressed. Despite these preliminary assumptions, the interview guides were intentionally developed to allow all participants to answer questions as freely as possible and with no major emphasis on issues that might have led interviewees into positively justifying these assumptions. In fact, the responses from the interviews proved that neither the sheer data size and volume nor the question of access and data rights could be identified as the crucial issues at hand. Moreover, the entirety of the system of autonomous labs, principal investigators' lab data protocols, individual data management practices of post-doctorates and graduate students, participants' awareness to whom one should turn in the event of a data management issue, a server or application processing problem, or crises with data analysis and/or locating lost data proved to be much more frequently cited as discernable issues.

The Central Facility

Based on the interview with the director of the central facility, the main data management issues do, however, have connection with data size, number of files, and overall data bulk that is often generated through some of its analyses. The Structural Biology Central Facility conducts image processing and data analysis services in three main types of scientific inquiry, arranged in ascending order of data requirements and data outcomes: X-Ray Crystallography, Nuclear Magnetic Resonance Imaging, and Molecular Dynamics simulations of protein structures. X-Ray Crystallography experiments produce data outcomes on the order of hundreds of megabytes, Nuclear Magnetic Resonance Imaging experiments produce data outcomes on the order of a few gigabytes to under a hundred gigabytes, and Molecular Dynamics simulations can produce data outcomes on the order of terabytes, with millions of files generated per simulation. It is the Molecular Dynamics simulations that are most computationally demanding and that require the most human-data interaction, so the data issues involving Molecular Dynamics simulations handled by the central facility became the primary focus of investigation in this study.

The Molecular Dynamics simulations not only produce a vast quantity of data and output files, the simulations, themselves are especially computationally demanding. Many simulations require the simultaneous use of 128 CPUs from the processing server and, depending upon the duration of the simulation being generated and the number of protein variants being simulated, can take a year or longer to run and to generate a final outcome. Again, this data outcome can be on the order of terabytes and can contain millions of files. While disk space and storage space on the University's mass storage system is not a limitation, migrating data on the scale produced by Molecular Dynamics simulations from directory to another is an extreme limitation that can tie up the Central Facility's computer resources to such a degree that other processing needs cannot be performed.

The central facility director cited numerous cases in which researchers who had been working on long-term Molecular Dynamics simulations were preparing to leave the University. In order for the Central Facility to retain the data that an exiting researcher has produced, the Central Facility must have access to the researcher's private, University-issued and ID-protected mass storage space so the Molecular Dynamics data can be transferred into the Central Facility director's University-issued and ID-protected mass storage space. When transferring data of this size that contains an astounding number of files and directories, the data transfer from one user's mass storage space to another user's mass storage space can take weeks, assuming there are no system outages or upgrades performed during the data transfer.

The Central Facility Director also cited there is no standardized format for organization and that there are no strict conventions for providing recommendations for organizational nomenclature to the directories and files generated from a Molecular Dynamics simulation. The director does, however, provide training and gives strong recommendations for file organization and for file naming. As the researchers are not under the direct charge of the Central Facility, these suggestions can only remain suggestions. The Central Facility director cannot be aware of each lab's technical infrastructure, the requirements or the suggestions proposed by each principal investigator, or the general work- and data- flows expected and set as precedent in each lab.

Principal Investigators

The interview responses from the faculty and principal investigators demonstrated a wide range of knowledge of data management practices, conventions, heuristics, and formally-documented guidelines for how the researchers in each lab should conduct experimental inquiry, retain, share, back-up, and deposit data. Furthermore, all but one principal investigator noted the importance of general consistency of technological infrastructure and devices among researchers in the lab. All of the principal investigators who did cite their desire for consistency in technological devices across their research staff provided pre-loaded laptop computers for their students that, at the time of issue, are functionally identical, yet uniquely identifiable: it is possible to determine a specific computer belongs to—or is for use by—a specific researcher.

Beyond the consistency of issuing equally-equipped laptops to new researchers, the responses from interviews with faculty and principal investigators showed a wide variance in other technological devices that each principal investigator provided or expected that each researcher use in the lab. For example, three principal investigators remarked they issues each research member in their labs a portable external hard-drive upon beginning research work with the lab. The reasoning behind issuance of the external hard drives is identical: the external hard drives are to be used as back-up and long-term storage devices that will be left in the possession of the faculty or principal investigator when the researcher departs the University. Although the external drives are intended for like use, the principal investigators issuing them have quite different expectations for how the external drives are to be handled and maintained.

For example, one principal investigator indicated that the external hard drives were to remain at each researcher's desk in the lab at all times and that the external drives were not to leave the premises. Furthermore, the researchers in this lab are expected to back up the data once a week from their experiments and from their other pertinent research work. Upon being asked whether this is checked and enforced, the principal investigator responded that it is at the discretion of each researcher to ensure that this weekly back-up task is performed. The principal investigator in this instance noted that there was no central repository into which all of the data from the external hard drives is then deposited as a central back-up of the entire data mass of the lab's researchers. Once each researcher leaves the University, the external hard drives become part of a physical library in the principal investigator's office, as if each drive were a silent volume or book without any identifying label or other indicator of the drive's contents.

In another principal investigator's lab in which each researcher is issued an external hard drive, the drives are free for the researchers to treat as their own property as they conduct their research. This means that the researchers may take the external drives home with them, leave them in the lab, or otherwise use them as they please. The principal investigator remarked that the lab's main interest lay in retaining and maintaining the data stored on the external drives and not in retaining the devices themselves once the researchers have left the University. The principal investigator would not elaborate on how the research data would be extracted, culled, and retained for potential further use. When asked to elaborate, the principal investigator cited that his researchers' work demanded a great deal of time to process and that the current body of research was to be viewed as longitudinally analyzed. The principal investigator remarked that because his lab's research often leads into novel directions and because generated research data is therefore qualitatively and quantitatively quite different than could have been assumed at the outset, it is impossible to anticipate data management requirements until all research and data generation have concluded. Furthermore, this principal investigator noted that the external hard drives would most likely be outdated and obsolete by the time the researchers had completed their time—usually three to five years—at the University.

In the final instance of a principal investigator issuing external hard drives to each member of the lab, the principal investigator described a more regimented, centrally-networked function of the external hard drives. The principal investigator noted that each hard drive is to be freely used by the researcher to whom it was assigned. However, the principal investigator also noted that the lab had hired a dedicated assistant who is charged with ensuring the data from each external drive is received from each researcher on a bi-weekly basis. This was only one of two labs in which the principal investigator specified there is a dedicated lab staff member who specifically collected and maintained data from the labs' researchers and who, furthermore, enforced practices of updating and managing the data assets of the lab as a whole.

While the external hard drives were only issued by a portion of the principal investigators interviewed, this does not mean only those principal investigators who issued external hard drives described an interest in long-term data storage for potential re-use within the lab or for the potential collaboration with other, perhaps, yet to be identified, groups. Purposes of these collaborations range from re-verification at a later date of raw data, secondary data analyses, and research methodology, to extension of data outcomes to other researchers, to alternative reanalysis of saved research outcomes. In fact, all principal investigators expressed a willingness to allow their prior research findings, along with the data that support these findings, to exist within an archive of sorts that could be freely accessed by others within the institution. The main issue in the shareability or the interoperability of this data lay in the form each principal investigator noted as a "best practice" for long-term data storage. Long-term storage strategies ranged from the aforementioned library of external hard drives to networked, freely-accessible storage that would be maintained either by the principal investigator or by a departmentally-appointed data manager, to physical archives of DVD's that would contain permanently-unalterable content.

Although the principal investigators have distinct preferences for how and where their labs' research data is held for long-term storage, none of the principal investigators interviewed for this study could quickly identify the location or the media corresponding to specific research or lab work. Furthermore, as all but one of the principal investigators have been faculty of the University for more than seven years, there are manifold storage locations and storage media that are not easily translated by or recalled by a single device or by a unified interface. As data management conventions and related data recording technologies have evolved, so, too, have the long-term data storage media and data recording practices changed for each principal investigator. All of the principal investigators cited numerous paper-format lab journals and work logs that contain data from experiments conducted as recently as five years ago. Principal investigators also cited the utility of written, paper-format lab journals for current research and all principal investigators noted that there are current post-doctorates and graduate students who still utilize some form of written, paper-format research journals or scratch pads that might or might not be transferred to the electronic data storage media that are to be left with the labs as research proceeds or once research is concluded.

Furthermore, all principal investigators noted that there is a great volume of data contained within written lab journals or in outdated forms, such as 5 ¼ " floppy diskettes. Principal investigators also cited data is often stored as obsolete file types that are no longer supported: for example, data might have been produced or analyzed using a computer application that is no supported and, therefore, the proprietary file format used by such an application cannot be opened by current computer systems. Some principal investigators noted there are a few independently developed web tools and software applications that can translate old proprietary file formats into data that are compatible with current applications. Even though there might be a means of translating a minority of files, this does not sidestep the difficulty of locating a device that can accept the media forms on which the data are stored. All principal investigators remarked that their busy schedules, driven primarily by the necessity to publish new research, do not afford the principal investigators the time or the resources to view their entire data stores as a whole. Therefore, it is impossible for the principal investigators to develop a strategy to consolidate their research data into a form that can be easily accessed via current electronic storage conventions.

Post-Doctorates and Graduate Students

Post-doctorates' and graduate students' individual data management practices add another layer of complexity to creating long-term data archives for the labs in which they perform their research. From the responses given during interviews, the post-doctorates and graduate students have individual preferences for conducting their work. This includes the times at which they send queues of files to be processed by the applications on the research server, the manner in which the post-doctorates and graduate students maintain the storage space allocated for them on the research server, and the regularity with which the post-doctorates and graduate students migrate their processed data out of the research server's scratch storage space to avoid deletion rules that have been established by the research computing division. Across the board, all post-doctorates and graduate students in the study responded they are given a brief introduction to the processing server and to how main workflows function; all post-doctorates and graduate students answered they are given no formal guidelines on scheduling their processing runs; all post-doctorates and graduate students interviewed also responded they are not provided with either heuristics or with recommendations for managing, organizing, or naming their data. Post-doctorate and graduate student respondents noted two main strategies for managing the data they generated during their research: one, data management practices learned from other projects or from prior general experience, and, two, advice given to them by fellow post-doctorates and/or graduate students performing similar research.

Despite the high variability and the informal nature of post-doctorates' and graduate students' data management tactics, researchers face common and specific infrastructural constraints. The Research Computing team who maintain the technological infrastructure has developed these constraints and the constraints are a set of general rules that hold true to all users unless user-specific alterations are requested and made in advance. There are deletion rules that are based on data and file latency, in which files that have not been accessed or moved are automatically deleted after twenty-one days. The research server does not provide a warning or a notification that the files are nearing, or have reached, the research server's limit. Therefore, the post-doctorates and graduate students who interact with the applications must remain aware of which files are stored in scratch storage, and how long the files have been latent. The research server also prohibits users from overreaching their maximum allowed storage space. Once a post-doctorate or a graduate student has reached the maximum allowable storage limit, all queues and processes related to that user are halted. Again, there is no warning or notification sent to the users of the research server that they are approaching or that they have reached the ceiling in terms of scratch storage space.

This interaction with the research server leaves the post-doctorates and graduate students, according to responses given during interviews, feeling they are remotely accessing a system that is unresponsive to their day-to-day data needs. Furthermore, the post-doctorates and the graduate students indicated they often forget to check the status of their scratch storage and they often lose data because they have reached the latency period or because their queues of molecular dynamics processing tasks have filled their scratch storage space. Because Molecular Dynamics simulations are dependent on a sequential series of files that indicate coordinates and parameters of protein molecules, any break in the sequence of output files can cause an interruption in the sequence of protein files that must be maintained. The sequence of files output by the applications on the research server can be compared to frames on a film: each file is directly spatially and temporally related to the file that has preceded it and each file is the foundation for calculating the spatial and the temporal characteristics of the file that follows it in sequence. Just as a moving reel of film provides persistence of vision to project moving images from still frames, the sequence of output files generated on the research server provides a longitudinal set of coordinates and properties that will ultimately be translated into a moving image of the protein being investigated.

The main issue from the perspectives of the post-doctorates and the graduate students is they are left on their own to migrate data from scratch storage into their individual university-ID assigned mass storage space. Interviewees all responded that there are no heuristics or reminders given in order to ensure the post-doctorates and graduate students remain mindful of the status of their data in scratch storage. Furthermore, since the files being generated are only a portion of the entire data set, all post-doctorates and graduate students reported they often save these data files to their laptop computers, to removable storage (such as USB thumb drives), or, where applicable, to their external hard drives. None of the post-doctorates or graduate students indicated they directly transfer their data from scratch storage to their university ID assigned mass storage space.   There were several reasons given by the interviewees for their intentional use of multiple storage devices to house portions of their molecular dynamics data. Some mentioned it is quicker and easier to capture the data from the scratch storage space and to transfer it to a laptop or to some form of external data storage. Those who cited the speed and the simplicity of storing the data onto external devices instead of directly migrating data from scratch storage to their private mass storage reported that they would later copy the files into their mass storage spaces as necessary and as the molecular dynamics files began to approach the storage limits of these external devices.

It is important to remember the Molecular Dynamics simulations can take months of continuous processing—sometimes over a year of continuous processing—before a successful longitudinal sequence of files is produced. Molecular dynamics simulations can often produce 2 terabytes of raw data, all of which is sequentially interdependent, before the data can be post-processed to generate a moving image of the protein being investigated. Therefore, because of the number of files generated and the time required to generate these files, it is likely there might be gaps within the sequence of files, there might be files that were unsuccessfully processed, there might be files that were deleted because they overran latency and scratch storage space constraints. When asked whether it was ever necessary to return to a prior portion of the sequence of output files because there was an irregularity or a break in the sequence of output data, all post-doctorates and graduate students responded they did have to sometimes return to a prior point in the processing sequence. Post-doctorates and graduate students noted they had to diligently locate a point from which they could confidently resume their research and generate reliable data.

Although the post-doctorates and graduate students as a whole had great confidence in how they managed their data, none of the post-doctorates or the graduate students indicated they used any shared conventions to name their files or to otherwise indicate specific file descriptors. Because the post-doctorates and the graduate students use idiosyncratic forms of generating file nomenclature that is based upon recommendations or conventions, it is often the case that files are temporarily lost. The terms "temporarily lost," indicate a different type of data loss than instances in which files have been deleted and they are no longer available. In instances of data being temporarily lost, the data was, in fact, intact and saved on the post-doctorates' or graduate students' storage devices. However, because there are no standards or heuristics for file nomenclature and for the organization of file structures, the data appeared lost. The data was, in fact, still available—the data was inaccessible. Interviewees responded these instances of lost or submerged data were not uncommon and that, usually, the post-doctorates and the graduate students would stumble across the data inadvertently at a later date.

Some post-doctorate and graduate students noted they would locate temporarily lost data before undergoing attempts to regenerate the data they had assumed they had lost. More frequently, however, post-doctorates and graduate students would find temporarily lost data after they had undertaken the painstaking task of returning to points in their research processes from which they could recreate the data they assumed they had lost. While locating temporarily lost data before a researcher undertakes the process of regenerating it might appear as less of an inconvenience to the researcher, both types of instances of temporarily losing data can be seen as significantly disruptive. Moreover, depending on the frequency with which instances of temporarily losing data might occur, temporary data losses can cumulatively significantly disrupt expectations of progress in a research progress. Interviewees also remarked temporarily losing data could sometimes create disruptions in the researchers' confidence in successfully completing their work.

Between Researchers and the Central Facility

Once the molecular Dynamics Data are generated and a run has concluded, the post-doctorates and the graduate students utilize the Central Facility to post-process the raw Molecular Dynamics data and to generate moving images of the protein structures being investigated. For the Central Facility, this can prove to be difficult because the post-doctorates and the graduate students approach the director of the Central Facility with data that is not named following any conventions, and that that might be stored across a combination of various devices and mass storage. The burden is now upon both parties: the researcher and the Central Facility's director, to make sense of which data is useful, to decide which files are necessary to process, to discern the order in which the files must be processed, and to evaluate which files must be saved and which data can be deleted.

Because the Central Facility has limited storage space, the director usually does not use the Central Facility's server to store the data from the molecular dynamics runs. There are instances in which the Central Facility director will transfer data from a researcher's mass storage space into the director's mass storage space. A unique user ID and password that are secure and prohibit group use or sharing of mass storage space individually protect university ID-based mass storage. In instances in which the director migrates data from a researcher’s mass storage space into the director’s own mass storage space, the researcher must grant access to the Central Facility’s director and the migration occurs as a transfer of files from one directory to another. Because of the number of files and the volume of data to be transferred, it can frequently take more than a week of continuous copying of files to migrate data from the researcher’s mass storage space into the Central Facility director’s mass storage space.

Because the data management tasks placed before the Central Facility’s director are unpredictable and often require imaginative strategies to best understand the entire meaning of a bulk of the Molecular Dynamics data, the Central Facility’s director is burdened with data management tasks that cannot be anticipated, that often require a great deal of investigation to make sense of how a researcher’s data is organized and named, and that often require that the researcher take several steps backward because some data must be re-generated. The Central Facility’s director is the only person capable of untangling and making sense of where to begin and how to find a rational strategy when handling data that needs to be reorganized, renamed, re-ordered, or otherwise brought to a stage in which the data are in a coherent state ready to proceed with post-processing.

The Research Computing Team

The members of the Research Computing Team that manages and maintains the research server are able to assist if there are errors that arise during the processing of

Molecular Dynamics simulations on the research server. For example, if a researcher cannot access the server or acquire enough CPUs to run a molecular dynamics queue, or if there is an issue with Molecular Dynamics runs failing for unapparent reasons, the research computing division can assist in troubleshooting the infrastructure, in identifying issues with processing tasks, and in recovering lost or corrupt files. When interviewed, members of the research computing division responded they receive very few requests from researchers to assist in handling server-based issues. One member of the research computing division remarked the research computing division is eager to help and to offer resources as needed; the difficulty is researchers do not approach the research computing division for assistance. It is unclear whether researchers are aware there are members of the Research Computing division available for certain troubleshooting assistance and researchers choose not to contact the research computing division or whether the researchers are unaware that there is a Research Computing division willing and able to offer timely assistance during certain data crises. Moreover, members of the Research Computing division are eager to offer assistance in affording researchers greater scratch storage space and extended latency periods before data are deleted—all the researchers need do is ask the Research Computing division. Members of the Research Computing division do not, however, possess domain-specific knowledge to aid in assessing the quality of generated data or to suggest best practices for performing Molecular Dynamics simulations.

Again, most of the strategies and the qualitative advice are given by the Central Facility’s director in the form of informal recommendations for file nomenclature and as strategies for file organization. Although the Central Facility’s director has developed these strategies from hands-on experience and the director can best advise how a researcher can tailor these strategies to best meet the researcher's objective, it is most often the case that researchers do not follow the Central Facility's director's recommendations. Respondents reported they did not follow the Central Facility's director's recommendations for two reasons: first, researchers felt their organizational strategies and structures were already sufficient and, second, researchers reported their organizational strategies and structures were too large, to complex, and/or too diverse to re-organize at the data analysis stage The Central Facility's director is faced with having to work with different operating systems—most labs are PC-based, one lab uses MacIntosh computers, and one lab runs a UNIX-based system. The diverse computer infrastructures and operating systems further complicate the issues researchers present to the Central Facility's director. Yet the post-doctorates and the graduate students repeatedly mentioned they turn to the Central Facility's director when there are failures in processing molecular dynamics simulations and when there are other research server issues, such as deleted or lost files, that would better be handled by the Research Computing division.

The Health Affairs Library

The Health Affairs Library has taken an interest in expanding the services it provides to the biomedical research community. In addition to providing access to electronic resources, information seeking and information retrieval strategies, the Health Affairs Library is deeply invested in investigating how it can provide useful data management assistance for the biomedical research community. The Central Facility is an opportune example of how the Health Affairs Library could begin to propose services that can aid the Central Facility and relieve much of the confusion and unpredictable obstacles. Currently, the Central Facility's director must successfully resolve most issues in order to make sense of and capably process data provided from manifold labs that use manifold computing platforms and that utilize manifold data management strategies. The Health Affairs Library intends to contextually investigate the system of tasks, data management issues, and data storage issues that the Central Facility currently faces. Although this is new territory for the Health Affairs Library, the Library's administration and staff are certain that providing data management services is viable and that it will offer a unique support framework that will reinvigorate the presence of the Health Affairs Library.

As there are few (if any) exemplars from which the Health Affairs Library can draw strategies or can adopt currently provided data management services, the Health Affairs Library is both relying on its own strengths within the current organizational structure and actively looking for campus collaborators with whom the Health Affairs Library can develop a course of action in developing data management services. All members of the Health Affairs Library who were interviewed for this study identified the Health Affairs Library already has a number of strengths that are particularly well suited to the library extending its support into new territory. First, and most, important, the Library's liaison structure is itself key to developing new services because the liaisons are continually tracking and reassessing current patrons' needs. The Health Affairs Library's liaisons are librarians who are assigned to provide library services and perform outreach support to the health science schools on campus. There are liaisons assigned to each of the major health sciences schools such as the School of Medicine, the School of Nursing, the School of Dentistry, the School of Pharmacy, and the School of Public Health. The liaisons function as members of the User Services department and provide myriad services that are domain-specific and that are general in scope.   For example, the liaisons are charged with general user services tasks such as circulation, reference desk, handling library chat and answering incoming calls to the Library's telephone reference service. The liaisons are therefore able to experience a diverse set of requests that are not necessarily linked with each liaison's domain assignment. The liaisons are also charged with remaining abreast of current practices and the liaisons are encouraged to conduct their own research that furthers librarianship. The Health Affairs Library's liaisons frequently conduct research in collaboration with the schools to which the liaisons provide services and the liaisons are encouraged to collaborate with faculty and students from the Information and Library Science to pursue salient and novel research. The Health Affairs Library proudly displays its extensive research background by displaying presented research posters and by holding presentations on current and published research.

The Library's current inquiry into providing data management services is yet another research opportunity for the library to be an exemplar. Within the university, the Health Affairs Library is a wellspring for seeking answers to emerging data management questions and members of the Health Affairs Library currently collaborate on a data management team consisting of librarians and library representatives from other university library units. The Health Affairs Library has been key to furthering the work of this interlibrary data management group and identifying biomedical research data management practices and services is one of the largest contributions to the ongoing agenda of the data management group. Two of the respondents from the Health Affairs Library directly mentioned the data management group as a significant peer resource group of campus collaborators.

Three members of the Health Affairs Library were interviewed for this study: a liaison who currently collaborates with biomedical researchers on campus and two members of the Health Affairs Library's management team. All three members of The Health Affairs Library also see the School of Information and Library Science as a strong partner to develop pathways that place librarians and information professionals into the necessary research contexts that will require data management support services. Affiliated with the School of Information and Library Science is a group of supercomputing specialists who are developing new strategies for creating stable archives that support biomedical data and that provide capabilities such as data repurposing and opportunities to engage campus research groups that could benefit from utilizing shared data. The Health Affairs Library has its own strengths in metadata assignment, cataloging, database development and database management, along with unique perspectives that view data differently than researchers in the biomedical community. There are many unknowns in reaching out to offer data management services because the services do not already exist—this is a great motivator for the Health Affairs Library to develop robust and practical services that can extend the Health Affairs Library's reach deeper into the biomedical research community.

The Health Affairs Library has a team of metadata experts who are committed to maintaining and to refining the Library's system of identifying, naming, and organizing information resources. Based on the findings of this study, the most frequent needs for the Structural Biology Central Facility's users are for assistance in the organization of information and in the creation of robust data organization schemes. An assumption at the outset of this study was that the Health Affairs Library might consider housing or managing the data itself. While e-Science discussions and the literature the literature on e-Science return to data volume, data storage resources, and computational infrastructure as primary to better handling e-Science needs, identifying the technological constraints of e-Science systems are only a beginning for further inquiry. The scope of e-Science researchers' needs reaches beyond technological and computational requirements and into how e-Science researchers can develop effective practices to work within the boundaries these technological requirements have constructed.

The Health Affairs Library's administration and staff are certain the Health Affairs Library's role will not be to house or store research data. The Health Affairs Library respondents noted the Library's role would most likely be to add levels of value and additional salience to current data stores. Respondents noted recommending data organization schemes, recommendations for metadata assignment, best practices for data management and storage, and recommendations for data archiving for data reuse as valuable examples of data management services. The Health Affairs Library also seeks to expand its current liaison structure, in which there are dedicated librarians affiliated with each of the schools in the biomedical-oriented portion of campus. Because the liaison structure already supports visibility of the librarians within specific academic contexts and because it affords librarians opportunities to be directly involved with research faculty in the research faculty's own environment. This proven structure is a robust foundation for preliminary engagement into new biomedical research environments.

Two respondents stated the Health Affairs Library does not currently have the data storage space or the computational resources to be handed en masse the data processing and the data storage tasks the Structural Biology Central Facility's users would require. All respondents noted the Health Affairs Library does, however, maintain a responsive in-house information technology group that supports the library's databases and that maintains the library's online presence. Moreover, one Health Affrais Library staff member noted the Research Computing division is already able to provide and to support the users' computational needs. There are, however, information, human, and organizational requirements that must be addressed no matter the technological requirements. Because the Health Affairs Library is founded on providing ongoing and responsive services to a substantial user clientele, the Health Affairs Library is in the best position to assess users' needs across disciplines and roles: this a service the Health Affairs Library is already providing through general services and through liaisons' outreach.

Furthermore, the interviewed members from the Health Affairs Library noted many of the support services the Library is provides for its own operation would benefit the Central Facility's data management. The Health Affairs Library engages in support services such as metadata generation, metadata assignment, recommendations for organizational structures best suited for research data, and recommendations for shared ontological resources that support researchers' data needs would be foremost to providing assistance to the Structural Biology Central Facility's users. Again, the Health Affairs Library is expert in providing these services to its own users and for its own purposes and the Health Affairs Library is adept at maintaining and flexibly revising these for its own resources while focusing on its patrons' needs.

The purpose of this study is to identify the parameters for a new service environment for which the Health Affairs Library could provide needed data management support. Creating new services might seem a tempting idea, especially because the e-Science literature posits great computational and technological infrastructural needs. However, seeking to extend the strengths of current Health Affairs Library services in the aid biomedical research domains was commonly identified as the best way to leverage the expertise that has aided the Health Affairs Library identify the Structural Biology Central Facility as a group in need. As with the Health Affairs Library's other outreach efforts, namely the extension of services through the library's liaisons, the members of the Health Affairs Library view this as an opportunity to extend already proven and established tactics of providing organizational support. Moreover, through a network of liaisons, the Health Affairs Library can pull together its existing services and augment them to best identify and address the needs of the Central Facility. Furthermore, the Health Affairs Library is a campus resource with many connections to other potential collaborators on campus, including the Research Computing Team, the School of Information and Library Science, science domain experts and communities of scientific research, and fellow members of the campus-wide data management community.

Analysis, Discussion, and Conclusion

From this study seven main issues were identified as the primary obstacles in the Central Facility’s data management:

1. Research data is frequently difficult to keep up with because of server infrastructural storage and deletion policies that constrain data storage during processing.
2. Research data is often stored across multiple devices and directory structures that are not necessarily part of the same general data storage architecture.
3. Research data is inconsistently named or it is stored within inconsistent file structures.
4. Users are unaware of support that can be provided to them from Research Computing at the point of research data generation.
5. Data is inconsistently stored for long-term use and is often difficult to retrieve because it is not findable or searchable.
6. Data is stored in formats, in systems, or on media that become obsolete and that ultimately prohibit access: timely updates and migration of data to new technologies.
7. Data is not easily shared with the Central Facility and data migration from researchers’ storage structures to the Central Facility’s data storage system is demanding in terms of technical resources and time.

To identify the best opportunities to provide recommendations for research data management support, one must remember most of the researchers in this study are pursuing research that is performed independently; researchers are not usually performing their routine research tasks in collaborative teams. Because most of the researchers are functioning in independent research environments it is important to remain respectful of the researchers’ methodologies they implement as they are conducting their research. The purpose of this study is not to confound or disrupt the domain-specific research needs of the researchers, but to augment their data management practices so researchers and the Central Facility benefit at each stage of their research and all who collaborate with the researchers can realize these benefits. As the Structural Biology Central Facility is the common actor in sharing data outcomes for secondary analysis, it is at the interface with the Central Facility that it seems most possible to aid the greatest number of participants. Furthermore, it is at the point of sharing the research outcomes with the Central Facility that the management of research data gains its most tangible collaborative dimension.

Although this study investigates the practices of research scientists performing domain-specific research within the biomedical sciences, it is important to remember the scope of this study is to investigate how the Health Affairs Library can contribute services to aid these researchers. In this way, the Health Affairs Library seeks to augment the researchers' practices and offer support drawn from the Library's set of skills and within the scope of services the Health Affairs Library would be able to provide. Just as these goals have been set and acknowledged by the Health Affairs Library, other members of the data management community researching into e-Science note stakeholders must be willing to acknowledge and refine the sets of domain-specific skills they can provide. More precisely, as e-Science is, by definition, collaborative, all parties involved in these collaborative forms of research contribute their expertise from various domains and from different perspectives (Hunsinger,2005; Pace, Bardzell, & Fox, 2010) . It is at these intersections of multidimensional disciplinary interaction that e-Science and has developed as a promising form of collaborative scientific research that and it is into these intersections that additional resources and contributions must be added.

As the Central Facility collaborates with numerous principal investigators who, in turn, each have teams of several post-doctorates and graduate students, the Central Facility is collaborating with a network of researchers who, even though they, at times, might be part of the same research group, are working primarily independently on individual projects at the time research data is generated. These individual research projects are most frequently not part of a larger team effort that involves coordinated data inputs from other researchers at the time the research data is being generated. Therefore, many of the individual post-doctorate and graduate researchers are primarily doing their work individually until it becomes necessary for them to share their research outcomes during later analysis. Because research is being conducted in solitude, most researchers have individualized strategies for processing and managing their data. This study showed only two examples of principal investigators who collaborated in developing intra-lab strategies for managing the data outcomes of individual research projects. Furthermore, these two labs articulated they have deliberate strategies for managing their labs' data. Upon further investigation, however, these strategies primarily concern maintaining data storage devices and research hardware and not developing deliberate strategies of how research data is managed. This demonstrates how the research labs queried in this study could benefit from assistance in developing practices to best manage how research data is created, how and when research data is manipulated and stored, how research data is named and tagged with identifiers, how research data is transferred from user to user and from division to division, how research data is archived for long-term storage, and how research data can be reused for future research.

The Research Computing team is very willing to hear from researchers during the research process. Research Computing team members want to be involved in understanding how researchers can get more out of the research server system. In order for the Research Computing team to better understand possible ways to modify and optimize the research server architecture, members of the Research Computing team would like to be involved in a dialogue with the users who interact with the research server. Through collaboration with research scientists, the Research Computing Team can gain an informed understanding of what is working, whether any obstacles arise that impede researchers performing their research, and whether the researchers have any suggestions on how to resolve these issues. Most frequently, the Research Computing team is not brought in to assess system issues until there has been a major issue and the issue has been resolved to a certain degree or the researchers have moved forward with their own strategies to compensate. The research computing server is, from the perspective of the Research Computing team who manage it, a resource that is available to a diverse community of users across campus and, as such, the research computing division is providing a set of services in much the same way as the Central Facility. Although only a handful of departments and disciplines make up the most frequent groups who utilize the server, the server is a campus-wide resource that is not necessarily affiliated with certain departments or research teams. In this sense, members of the Research Computing team want to maintain their openness to collaborate with campus stakeholders to better the ways the research computing resources might be better utilized overall.

Because the principal investigators and the lab researchers are independently assessing and addressing their data needs, the issues each lab experiences are often transferred to the Central Facility only when the Central Facility is brought on board to assist during secondary data analysis. These data management issues are often difficult for the Central Facility to anticipate, and the Central Facility is not brought in to collaborate with the researchers before these data management issues arise. In many cases, the Central Facility must make sense of data nomenclature and data storage schemes once these storage conventions have solidified well into the course of the research. Members of the Research Computing team are available to address technical issues during the course of data production but the Research Computing team members are primarily responsible for maintaining the applications through which the research data is produced and the Research Computing team members do not possess research domain specific knowledge. Additionally, one of the largest questions at hand is how the Central Facility can maintain continued access to data once a researcher has left the university and how unexpected changes in research members can have a minimal impact on how the Central Facility provides its services.

The Health Affairs Library can be an active participant in assisting biomedical researchers in their data management. This study identifies numerous avenues the Health Affairs Library can pursue to develop a strong presence in assisting the Molecular Biology Central Facility. E-science has been identified as a new frontier for academic scientific research and much of the literature focuses on the technical requirements of e-Science research systems and the technological demands faced by researchers engaged in the practice. What is often left out of the discussion is how to manage the framework of human data management practices that are both part and parcel of e-Science research (Hunsinger, 2005; Branco & Moreau, 2006;Pace et al., 2010; Kowalczyk, 2011). The Health Affairs Library can effectively engage in offering data management support to e-Science researchers because many of these human factors are already being addressed in similar fashion as the Health Affairs Library currently supports its patrons. The

Structural Biology Central Facility would be a new environment into which the Health Affairs Library could extend its current support services and through which the Health Affairs Library can continue to refine the services it provides. Currently, the population of biomedical researchers interviewed in this study are underutilizing the Health Affairs Library as a collaborative partner and the biomedical research communities identified in this study are individually attempting to work through their data management issues.

Exactly how the Health Affairs Library can demonstrate a visible role to biomedical researchers remains one of the largest issues and, just as e-Science has developed through complex connections among domain researchers that longitudinally identify the requirements and applications of shared research data, the Health Affairs Library must position itself as a long-term partner and collaborator in e-Science. In this way, the Health Affairs Library is just beginning its involvement and must understand that much of the challenges identified cannot be effectively solved by only paying attention to the short term. Moreover, as e-Science will continue to develop, the Health Affairs Library must integrate itself into the biomedical research environments through practices that will identify opportunities and that will illuminate flexible and integrative strategies. Individual research teams might muddle through the process and find a solution they believe best suits them, only to find they have missed some large concepts that information professionals might better be able to advise on. Furthermore, the Library's research is founded on pressing information organization topics and is interested in addressing the information needs an in researching the state of the art of information management, and of providing access to information resources.

The Health Affairs Library must look at the opportunities to engage in e-Science data management support as opportunities to ask questions and be informed by the research community, not as opportunities to merely implement suggested practices and solutions. Therefore, this initial phase for the Health Affairs Library should focus on an environmental assessment of current stakeholders; the Health Affairs Library must acknowledge potential answers will come from continually returning to and reassessing answers provided to initial questions. As many researchers in the fields of Information Science and Library Science have expressed, the main objectives at this stage are on sensitive and engaged assessment and not on issuing groundbreaking recommendations (Lyon, 2009; Lewis, 2010; Corrall, 2012; Lyon, 2012). Furthermore, many information professionals researching e-Science note it is imperative for libraries to understand exactly how best they can participate in a collaborative way without overextending themselves too far outside of their own expert domains (Hey & Hey, 2006; Gold, 2007, Lewis, 2010). Therefore, this study has identified a number of recommendations that the Health Affairs Library can implement in order to become a sensitive agent of change that contributes its user-centered imprint and momentum to e-Science research.

One of the Health Affairs Library's services is as a central meeting place for members of the campus community. Therefore, the Health Affairs Library should begin by offering collaborative meetings with research teams to see examples of current data production and data management workflows. From these meetings, the Health Affairs Library will be able to assess current limitations and the need for further investigative research and training. A focus that should be underlined in these collaborative meetings should be evaluating the current sets of diverse storage media and storage locations to understand technological and human limitations in managing data. The Health Affairs Library should also act as key collaborator in allowing the Research Computing division to participate in group discussion and assessment of technological needs. Bringing in members of the Research Computing division would allow the Health Affairs Library to recommend campus strategies for new forms of collaborative storage. The Health Affairs Library can also advocate for smart technologies and for innovations in data management solutions that can be provided to groups campus-wide. The Health Affairs Library already exists as an innovator that centrally houses technologies for collaborative learning and for innovative data visualization. This is an opportunity for the Health Affairs Library to expose more potential patrons to the Library's services and for the Library to continue with its own development of ground breaking user services. The Health Affairs Library can provide shared spaces for learning, for collaborative meeting, and for open presentation of findings.

Many of the data management difficulties identified in e-Science research are not directly linked with specific devices or computing platforms, however (Atkins, 2003; Newman, 2003; Gold, 2007; Antunes, Baraterio, Cabral, Borbhina, & Rodriguez, 2009; Bietz & Lee, 2012). Several difficulties noted by the researchers and the Central Facility in this study are based on difficulties in generating data organizational schemes and data nomenclature. The Health Affairs Library is knowledgeable in understanding the best ways to create networks and systems of cataloging, naming, storing, and protecting digital information assets so there are persistent records of what something is, when it was created, who created it, when it was updated, how it might be of reference to another discipline, how best to keep it, and how best to keep it so it is permanently shareable. It is imperative the Health Affairs Library apply its own knowledge in cataloging, metadata assignment, and management of electronic resources to assess the effectiveness of the Central Facility collaborators' storage structures and data organizational strategies to determine whether or not there are better ways of naming and organizing files. The Health Affairs Library should leverage its experience in organizing its own collections of print and electronic resources as a set of best practices that can be handed to the principal investigators as they bring new post-doctorates and graduate students into their research teams. These best practices should be developed in collaboration with the labs' needs and should be refined frequently to ensure the best practices remain current and pertinent.

Liz Lyon has introduced the practice of performing a robust data audit for identifying the key issues in research data management and in assessing the best ways to implement measurable and deliberate progress in research data management (Lyon, 2007; Lyon, Coles, Duke, & Koch, 2008; Lyon, 2009; Lyon, 2010; Lyon, 2012). The Data Audit Framework consists of a number of interdependent actors who must provide expert input in order to gain a rich picture of the current state of research data and how to successfully move forward in ensuring the entire data management environment is assessed and reassessed as necessary. The Health Affairs Library is currently moving forward with the preliminary investigative steps in creating the necessary campus relationships in developing a Data Audit Framework and further identifying and inviting key campus partners into the fold is imperative. Performing an initial data audit might involve a large group of collaborators, but the ongoing efforts to reassess biomedical research data will likely not require all collaborators to participate all every phase of a data audit. Therefore, the Health Affairs Library should develop training resources for researchers to perform self-checks and assessments as the researchers are generating their research data.

Several key partners are already collaborating with the Health Affairs Library and these relationships must be seen as central to the collaboration necessitated by e-Science. The Health Affairs Library must continue to partner with the School of Information and Library Science to investigate new technologies from the field and to bring a network of research-oriented academicians into the discussion. Furthermore, the Health Affairs Library must continue to leverage its liaison structure to identify other groups of researchers who might be experiencing similar issues or who might offer insight from how they have handled similar issues. The Library's liaisons can propose initiatives to identify hidden populations that might be suffering from the same issues as the Central Facility. Once these communities of researchers in need of data management support have been identified, the liaisons can refer them through a network of campus experts in order to find the best data management partners. The liaisons should further embed themselves into the schools they assist so they are increasing their visibility and their perceived value. For example, the liaisons could develop specific on-site data management training events that would cater specifically to the university schools the liaisons serve. Therefore the Library's best information organization skills can be applied to different academic domains and in different environments.

Members of the biomedical research community interviewed for this study remarked the NSF and the NIH have begun requiring data management protocols to be included into the grant writing process. As this study shows, the principal investigators do not currently have a set of institutional or disciplinary data management guidelines they can efficiently incorporate into their grants. This is a promising opportunity for the Health Affairs Library to develop data management resources for researchers to simplify their grant writing. Furthermore, specific recommendations for data management could be revisited as principal investigators proceed with their research and knowledge gleaned from the revision process could be made available to others. Future researchers would gain assistance in drafting management provisions, as they would be able to find reliable sets of proven recommendations. The Health Affairs Library is already flexible and accommodating to its patrons and this is an opportunity for the Health Affairs Library to contribute within a community of experts to offer new forms of support. Individual researchers would not left to suffer on their own and compromise as the issues arise; researchers could communicate with members of a community of expertise who can provide flexible solutions to aid researchers in solving their problems and can offer advice found from experience with other research disciplines.

As the Health Affairs Library progresses in its participation as a way to synthesize the efforts of key campus partners, the Library can recommend new systems of data devices, networks of storage and of connected technologies, and networks of communities of practice that will eventually be working in a collaborative manner once connections between research data and research methodologies are identified. The Health Affairs Library can serve as a steward for ensuring access and meaning are preserved. Furthermore, the Health Affairs Library is not a single lab or a single department within a discipline. The Health Affairs Library is a research division in its own right with connections to departments on the campus it serves and, through its membership in professional organizations and in the publication of its research, the

Health Affairs Library maintains visibility and pursues to discuss and share its own findings with peer institutions. Therefore, the Health Affairs Library can be seen as a key player in promoting collaboration with divisions at other research institutions. From these potential connections and likely opportunities for sharing research, the Health Affairs Library can be a progressive agent for not only changing the research data management practices of the research divisions on the same campus, but would serve as an advocate of furthering data management practiced through inter-institutional cross-pollenization. The Health Affairs Library can aid in transforming the research experiences of small groups who currently suffer independently as they attempt to find their own solutions and best practices.

To summarize, there are two main campus stakeholders identified by this study. Although there is a wealth of information provided by the post-doctorates and graduate students, the principal investigators, and the research computing division, one primary stakeholder is the Structural Biology Central Facility. The Central Facility is the entity that is exposed to diverse and, often unexpected, data management tasks that often must be handled on the fly. The Central Facility also deals with the greatest number of agents: principal investigators, post-doctorates and graduate students, research computing specialists, and so on. The Health Affairs Library is the second primary stakeholder because the Health Affairs Library seeks to leverage its own abilities and strengths along with the strengths of potential campus collaborators to create practicable data management services to the biomedical research community—the Central Facility is its first opportunity.

The Health Affairs Library is a division that serves a diverse campus-wide community of biomedical researchers. In this way, the Health Affairs Library is, as the Central Facility is, seeking to provide a visible set of resources that are sensitive not only to the users who call upon the Health Affairs Library for assistance, but likely pertinent to future users who might come to the Health Affairs Library with information and data management needs that are currently unknown or unidentifiable. And it is precisely because the Central Facility has begun its collaboration with the Health Affairs Library that proves the Health Affairs Library is present and visible as a valuable campus partner. Several researchers have identified the challenges of e-Science echo many of the electronic collections and resources management issues libraries must handle. These issues include but are not limited to collection management, electronic resource access management, cataloging, metadata development, metadata assignment, maintaining provenance of items in collections, mitigating access issues and anticipating obsolescence.

From this perspective, it is possible to see the Central Facility's data management issues as opportunities to extend current library practices and apply them into a new environment. All stakeholders in this study are especially open to the Health Affairs Library providing consultation and assistance in developing new data management strategies. Because the data management challenges are especially apparent at points in which researchers and the Central Facility collaborate, all parties are willing to invite an outside authority to assist because the current data management issues impede overall progress. There were no reservations or concerns from the researchers and the Central Facility that the Health Affairs Library's assistance would be intrusive or disruptive.

It is vital one considers the obstacles when proposing any kind of data management service. Issues of data organization, storage space, and technological infrastructure are obvious and have been identified here. As e-Science continues to develop and provide many more networked sources of research data, the issues identified here will undoubtedly expand and become more complex. These increasing complications have already been experienced as researchers have transitioned away from static forms of research that were not developed with connectivity and sharing of data as necessities. Technological limitations and obsolescence will inevitably be major concerns as many of the principal investigators already have volumes and volumes of data that are not available to anyone because these volumes are hand-written and on bookshelves, they are contained on obsolete data storage devices that cannot be accessed, some are in file formats that are no longer supported or that are no longer translatable into currently supported formats.

Numerous questions arise in how to address matters of obsolescence before it becomes a pressing likelihood. First, how does one comprehend obsolescence when future processing devices, data storage devices, file formats that don’t yet exist will make current technologies and devices appear obsolete? Also, how does one anticipate the future of biomedical disciplines and the development of new methods of research and inquiry that might surpass the utility of current data even if it were archived, readily available, and easily accessed? Are we looking at this challenge with wide eyes and the belief that technology holds most of the answers? Are there simple and more mundane practices or practical arrangements that might be small but that could be more robust over the long term?

Secondly, who or what is the best agent of change to develop data management practices? Is a library and are librarians and information professionals equipped with the methodologies, skills, and perspectives to provide services that offer utility and that provide results? Domain knowledge is imperative to comprehending the data management issues that currently exist and that must be anticipated within specific research domains. To this end, it is important for libraries, librarians, and information professionals to being exposing themselves to scientific domains and to gain training in understanding the science behind the types of research being conducted. The objective is not for the information professionals to become seasoned scientific researchers but for the information professionals to become contextually informed key collaborators. Furthermore, as the research and data management practices within research groups are already experiencing its own pressures, it is imperative for organizational, cultural, and practical boundaries be acknowledged and respected. Additionally, information professionals must frequently reach out across domains within its peers to determine others who are researching data management within different disciplines. It is likely that through this reaching out all involved will identify proven methodologies, key successes and failures, proposed systems, and sets of best practices.

Thirdly, how can e-Science data management create initiatives for institutions to curate and take stock of its institutional data assets? It is fundamental for central institutional divisions to prove their impartial abilities at being essentially linked to institution-wide objectives. As e-Science becomes more involved across research disciplines, institutions will be well served to invest collaborative effort into developing and maintaining repositories of knowledge that will not only serve the current e-Science research community but will extend knowledge into novel applications of further research. In this way, academic research institutions would fulfill their promises of interdepartmental academic collaboration and develop large stores of shareable data that can be applied in environments that cannot yet be imagined.

Finally, how can e-Science data management practices and recommendations be reincorporated into teaching and training within the disciplines of Information Science and Library Science? Information professionals must remain actively engaged in collaborative efforts to develop the next generations of e-Science systems and the future practitioners who will be involved in envisioning, implementing, and managing e-Science systems of the future. It is necessary for e-Science collaboration to occur through the development of curriculum that can be augmented with practical field experiences that expose current Information Science and Library Science students to embedded communities of biomedical research practice. Furthermore, the less visible agents, such as the Research Computing division, and other institutional groups that do not receive a great deal of interaction from research communities would be able to participate in low-risk endeavors that would require a lower level of investment and distribution of institutional resources, illuminating potential eagerness to contribute. As these opportunities for exposure to new e-Science research environments continue to develop, curricular opportunities for disciplines mutually influencing each other will bring needed longitudinal assistance to populations suffering through the pain of managing complicated systems of e-Science research data.

References

Antunes, G., Barateiro, J., Cabral, M., Borbinha, J., & Rodriguez, R. (2009). Preserving digital data in heterogeneous environments. *JCDL '09: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 345-348.

ARL—Association of Research Libraries (2007). Agenda for developing e-science in research libraries: ARL joint task force on library support for e-science final report & recommendations. Retrieved from http://www.arl.org/storage/documents/publications/arl-escience-agenda-nov07.pdf

Atkins, D. Droegemeier, K., Feldman, S., Garcia-Molina, H., Klein, M., Messerschmitt, D., Messina, P., Ostriker, J., Wright, M. (2003). Revolutionizing science and engineering through cyberinfrastructure: Report of the national science foundation blue-ribbon advisory panel on cyberinfrastructure. Retrieved from http://www.nsf.gov/cise/sci/reports/atkins.pdf

Barga, R., Fay, D., Guo, D., Newhouse, S., Simmhan, Y., & Szalay, A. (2008). CLADE '08: Proceedings of the 6th international workshop on challenges of large applications in distributed environments, pp. 63-67.

Barateiro, J., Borbinha, J., Antunes, G., & Freitas, F. (2009). Challenges on preserving scientific data with data grids. DaGreS '09: *Proceedings of the 1st ACM workshop on data grids for e-science*, pp. 17-22.

Bietz, M., Ferro, T., & Lee, C. (2012). Sustaining the development of cyberinfrastructure: An organization adapting to change. *CSCW '12: Proceedings of the ACM 2012 conference on computer supported cooperative work*, pp. 901-910.

Bietz, M. & Lee, C. (2012). Adapting cyberinfrastructure to new ccience: Tensions and strategies. *iConference '12: Proceedings of the 2012 iconference,* pp. 183-190.

Branco, M. & Moreau, L. (2006). Enabling provenance on large scale e-science applications. *Provenance and annotation of data*, Vol. 4145, pp. 55-63.

Corrall, S. (2012). Roles and responsibilities: Libraries, librarians and data. In G. Pryor (Ed.), *Managing research data* (pp 105-133). London: Facet.

Gold, A. (2007). Cyberinfrastructure, data, and libraries, Part 2 libraries and the data challenge: Roles and actions for libraries. D-Lib Magazine 13(9/10). Retrieved from http://www.dlib.org/dlib/september07/gold/09gold-pt2.html

Hedges, M., Hasan, A., & Blanke, T. (2007). Management and preservation of research data with iRODS. *CIMS '07: Proceedings of the ACM first workshop on cyberinfrastructure: information management in eScience*, pp 17-22.

Hey, T., & Hey, J. (2006). E-science and its implications for the library community. *Library hi tech*, 24(4), pp. 515-528.

Hey, T. & Trefethen, A. (2003). The data deluge: An e-science perspective. *Grid computing: making the global infrastructure a reality*, pp. 809-824.

Hunsinger, J. (2005). Reflexivity in e-science: Virtual communities and research institutions. SIGGROUP Bulletin, 25 (2), pp. 38-42.

Jankowski, N. (2007). Exploring e-science: An introduction. *Journal of Computer-Mediated Communication*, 12 (2), pp. 549-562.

Karasti, H., Baker, K. S., & Halkola E. (2006). Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (LTER) network. *Computer supported cooperative work*, 15, pp. 321-358.

Kowalczyk, S. (2011). Towards a model of the e-science data environment. *JCDL '11: Proceeding of the 11th annual international ACM/IEEE joint conference on digital libraries*, pp. 399-400.

Lyon, L. (2007). Dealing with data: Roles, rights, responsibilities, and relationships. UKOLN. Retrieved from http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_dealing_with_data.aspx\

Lyon, L., Coles, S., Duke, M., & Koch, T. (2008) Scaling up: Towards a federation of crystallography data repositories. UKOLN. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html#2008-05-12

Lyon, L. (2009). Open science at web-scale: Optimising participation and predictive potential. UKOLN. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html#november-2009

Lyon, L. (2010). Open science in the data decade. *Public service review central government edition*, 20, p. 145.

Lyon, L. (2012). The informatics transform: Re-engineering libraries for the data decade. *International journal of digital curation*, 7(1), pp. 126-138.

Nam, D., Lee, J., Hwang, S., Suh, Y., & Kim, B. (2008). Research process support with organizational flow in e-science. *7th IEEE/ACIS international conference on computer and information science*, pp. 655-660.

Newman, H., Ellisman, M., & Orcutt, J. (2003). Data-intensive e-science Frontier research. *Communications of the ACM,* 46 (11), pp. 68-77.

Pace, T., Bardzell, S., & Fox, G. (2010). Practice-centered e-science: A Practice turn perspective on cyberinfrastructure design. *GROUP '10: Proceedings of the 16th ACM International Conference on Supporting Group Work*, pp. 293-302.

Stewart, C., Simms, S., Plale, B., Link, M., Hancock, D., & Fox, G. (2010). What is cyberinfrastructure? *SIGUCCS '10: Proceedings of the 38th annual fall conference on SIGUCCS*, pp. 37-44.

Appendix

Profiles and Interview Questions of Subject Population Subgroups Investigating a Data Management Environment for Structural Bioinformatics Research Data

Central Facility Director

The Structural Biology Central Facility Director assists graduate students, post-doctoral scientists, and principal investigators in structural bioinformatics computation and visualization required for grant funded research publications. The main purpose of the Central Facility is to assist researcher's university-wide in incorporating structural biology/bioinformatics into their grants and publications. The Central Facility is there for researchers who don't have a traditional expertise in Structural Biology. Central Facility staff provide limited and temporary storage space for data inputs and outputs for these structural bioinformatics computation and visualizations including NMR spectroscopy, X-Ray crystallography, and molecular dynamics data. This storage space supplements student, post-doctorate and scientist individual network storage space on campus ITS systems. The Central Facility operates as an independent centralized service for processing structural bioinformatics data for members of the campus biomedical research community. As such, the Central Facility has no direct affiliation with any campus department. Furthermore, the Central Facility Director does not manage, oversee, or direct any employee outside of the Central Facility. All Researchers, Graduate Students, and IT & Resource Support Staff operate independently of the Central Facility and either provide services to or request services from the Central Facility.

- How many researchers do you work with?
- How many projects do you typically have going at one time?
- What types of analysis does your facility perform?
- What applications do these techniques require?
- How much computing time do these techniques require?
- Who provides computing support for the applications that you use?
- Where is the analysis performed? Is it performed on local hardware or on servers?
- What data formats do these applications require/use/produce?
- How much data is produced for a typical project?
- Where is this data stored?
- How much local storage space do you have?

- How much storage space is available to you?
- Who provides storage space and who manages it?
- Who has access to the data? How is this access controlled?
- Who has rights to the data?
- How long do you store data?
- Do you have any rules or guidelines for organizing or naming data?
- How are you involved in handling data for each researcher?
- Does any of the data get re-used for multiple analyses?
- Are there analytical procedures that you repeatedly use?
- Are there researchers that you work with on a repeated basis?
- What data, if any, needs to be retained and how long do you need to retain it?
- Are you interested in keeping any data in long-term or permanent storage?

Laboratory Scientists who Utilize Central Facility Services

These are the principal investigators and staff researchers who utilize the Structural Bioinformatics Central Facility services. These faculty and staff are currently primarily connected with the following departments, centers and programs: Biology, Chemistry, Biochemistry/Biophysics, Pharmacology, Cell & Developmental Biology, Cancer Center, and the Program in Molecular Biology and Biotechnology. Currently, 75% of people who use the Central Facility are affiliated with the Cancer Center. Principal investigators need persistent access to data associated with their research and publications. Although the principal investigators utilize the Structural Biology Central Facility's services, they are not under management of or in charge of any employee of the Central Facility. The Central Facility functions as a centralized service that is open to all departments that would require use of its services. There are no direct professional or organizational appointments or personnel positions that exist between the Central Facility and any members of the principal investigators subgroup.

- What types of analysis do you require for your work?
- Who generates data for these?
- What formats are these data in?
- How large is a typical dataset?
- Who analyzes this data?
- How long does a typical research project last?
- How many people are involved in creating/using the data?
- Where is the data stored?
- Do you have local data storage?
- What is your data storage capacity?
- Are you interested in long-term data storage or archiving your data?
- Who has access to the data?
- Who has rights to the data?
- Do you ever share data with other researchers?
- How do you transfer data from place to place or from person to person?
- Who provides you with technical support or assistance?
- Do you have any rules or conventions for organizing or naming your data? If so, what are they and how are they communicated or enforced?
- What data, if any, needs to be retained and how long do you need to retain it?
- Are you interested in keeping any data in long-term or permanent storage?

Graduate Students/Post-Doctorates Who Utilize Central Facility Services

Graduate students and post-doctoral fellows working on research teams with laboratory scientists are often the primary users of the Structural Bioinformatics Central Facility's systems. The Central Facility Director assists and guides them students with their work for the larger research project. As the members of the research team often conducting the structural bioinformatics computation and visualization, they also generate and manage the data associated with this work. Often this data is stored in their personal ITS systems space on the campus network which is campus password authenticated. Graduate students and post-doctoral fellows regularly leave the research team and institution when their programs are complete. Data they have been responsible for often needs to be transferred to others in the research team in order for it to be accessible to them. This data and related files vary in how well organized and identifiable they are. As with the Central Facility Affiliated Laboratory Scientists, graduate students and post-doctorates who utilize the Central Facility's services are not directly affiliated with the Central Facility itself. There are no direct professional or organizational appointments or personnel positions that exist between the Central Facility and any member of this subgroup.

- What types of data do you produce?
- What formats are these data in?
- Which applications do you use?
- How do you access these applications?
- Where do you store your data?
- What types of analysis do you perform on your data?
- Do you use different storage for different types of data?
- Do you follow any organization scheme or use any naming conventions to organize your data?
- Who has rights to the data?
- Do you share access to data with others?
- Who provides you with technical support, assistance, or training for using the data?
- Would it be possible for others to make use of your data for their own analysis?
- What data, if any, needs to be retained and how long do you need to retain it?
- Are you interested in keeping any data in long-term or permanent storage?

Research Computing and Resource Support Staff Who Provide Support to the Central Facility

Structural Biology Central Facility data storage and systems administration needs are handled by the combined efforts of a local Central Facility server/systems administrator, IT staff from the Center for Bioinformatics, and main campus ITS who manage campus network space of individual faculty, staff, and students. Cenrtral Facility IT and Resource Support Staff are not directly affiliated with the Central Facility. There are no direct professional or organizational appointments between the Central Facility or personnel positions that exist between the Central Facility and any member of this subgroup.

- What applications do you provide for the Central facility?
- Where are these applications located and how do they function?
- What data formats do these applications use/produce?
- How much data is produced by a typical analysis?
- Where is this data stored?
- Who has access to the data?
- Who has rights to the data?
- How much storage space is available?
- Is long-term storage space available?
- What support do you provide for the Central Facility?
- Do you provide guidelines for organizing or naming the data?
- What data, if any, needs to be retained and how long do you need to retain it?
- Are you interested in keeping any data in long-term or permanent storage?

Health Affairs Library Staff and Administration

Health Affairs Library leadership and staff supporting bioinformatics are interested in identifying new library services to support researchers and their growing data management needs. They are investigating new roles that information professions can best fill in support of e-science. The Director of the Structural Biology Central Facility invited library staff to work with her on the facility's data management challenges, particularly improving the organization, management, and transfer of data that is generated by graduate students and post-doctoral fellows and needs to be readily accessible to their principal investigators and research team colleagues.

- What role do you want to play in data management and what services would you provide?
- What are your objectives for including these new services into the library's current services?
- Do you have data storage or processing facilities?
- Are you interested in hosting the data?
- Are there any organizational models that you intend to follow with the Health Affairs Library's data management projects?
- Are there any information organization models, frameworks, or systems that you have in mind for assisting in data management?
- Who would be responsible for overseeing data management projects?
- How would the library administer these projects?
- What types of data are you interested in helping to manage?