In November 2007, the Joint Task Force on Library Support for E-Science from ARL issued a report entitled the “Agenda for Developing E-Science in Research Libraries.” This report outlined critical areas for engagement, as well as strategies for libraries to undertake in order to improve support for the E-Sciences at an institutional level. This paper examines five on-campus laboratories from UNC-Chapel Hill, and using the ARL guidelines and agendas projects how the UNC-CH library system can begin to improve support for E-Sciences.

Headings:

College and university libraries – Information Networks – Virtual Library – Research and the library
EXPANDING THE GRID: COLLABORATIVE OPPORTUNITIES IN E-SCIENCE

by
Christopher M. Granatino

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 Library Science.

Chapel Hill, North Carolina
July 2008

Approved by

_______________________________________
Barbara B. Moran
Table of Contents

Introduction ......................................................................................................................... 2
Review of Literature ........................................................................................................... 9
Methodology ....................................................................................................................... 13
Analysis ............................................................................................................................ 19
Conclusions ...................................................................................................................... 27
Bibliography ..................................................................................................................... 33
Appendix A: Interview Questionnaire .............................................................................. 35
Introduction

The term “e-Science” was first coined by John Taylor, serving at the time as the Director General of Research Councils at the Office of Science and Technology (OST) in the United Kingdom. Taylor stated, “e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it,” (Taylor 2001). In essence Taylor broadly defined e-Science as the supporting structures that allow for collaboration between researchers. Perceived from different angles, this definition can provoke a great deal of speculation regarding the evolving infrastructures for collaboration supported by new tools and technologies (Hey 2006). Since the definition of e-Science begins with global collaboration it can be difficult to determine exactly where to begin when developing support structures on institutional levels. Roles are still being determined by actors on all levels from government agencies to academic institutions to software and publishing industries (Gold 2007). As such, librarians are caught in a position where it may still be unclear how best to support their individual institutional infrastructures.

To assist with this problem, the Association of Research Libraries (ARL) Steering Committee for Scholarly Communication and for Research, Teaching, and Learning created a Joint Task Force on Library Support for E-Science in 2006. ARL is a nonprofit organization of 123 research libraries at comprehensive, research-extensive institutions in the US and Canada that share similar research missions, aspirations, and achievements (ARL website). The charge of this committee was to “raise awareness and position
research libraries to be players in this new arena.” (Lougee 2007). The five areas the Task Force were given to focus on were:

- Developing an understanding within the research library community about the issues and needs associated with e-science and cyberinfrastructure and the associated needs of scientists and researchers.
- Recommending approaches to addressing issues related to the curation of long-lived digital data, including the handling of simulations and storage of massive data sets.
- Engaging ARL members in the development of new roles for libraries as e-science infrastructure and services emerge at research institutions.
- Identify the skills needed as information professionals move into the emerging e-science landscape and encourage the development of information professionals prepared to assume new roles.
- Identify opportunities and recommend strategies for developing relationships with various government science agencies and other stakeholders such as scientific societies.

In November 2007, this Joint Committee released a report entitled “Agenda for Developing E-Science in Research Libraries”, which addresses key areas for concern among academic librarians, and strategies for improving support on institutional levels. In their report, the Joint Task Force identified three ‘critical areas for research library engagement’ and five sets of desired outcomes for research libraries to support e-sciences. The three critical areas are defined as:

- Data Issues and New Genres of Scholarly Communication
Virtual Organizations

Policy Development

The first area, data issues, refers to the role libraries must play in managing and preserving data (Lougee 2007). With the voluminous amount of data that is transferred among scientists, there needs to be support in place to assist in the handling and preservation of these data sets. This will require the further development of metadata standards to make these data sets usable, searchable, and stable over time. Anne Gold, a member of the Joint Committee elaborated in a paper from late 2007, stating,

A major effort in the arena of curation is to develop metadata practices and standards that will do several important things: first, they must make data understandable by computers; second, they must support discovery across heterogeneous data collections; and third, they must manage all of this across data scales from the small to the immense.

This issue of the knowledge discovery process is at the very core of e-science and cyberinfrastructure, as its acquisition should be the ideal end to which this extensive data sharing aspires. The challenge lies in the process of transforming large amounts of raw data into practical information that other parties can search and utilize. The nature of collaborative science relies upon the ability of the researcher to share, the ability to provide more access and reuse of intermediate material (Wright 2007). With tools such as RDF (Gold 2007), libraries can lend their expertise to researchers and support e-science by improving the preservation capabilities for data sets.

Virtual organizations, the second area, refer to online environments that bring distributed researchers together and provide relevant content, data, tools, and services to
assist collaborative work (Lougee 2007). To this end, the ARL report contends that
discipline-focused repositories could provide an important, core component for the
virtual organization to which relevant social tools for collaboration could be added.
Digital libraries and repositories are familiar territory to academic libraries, and examples
of collaborative workspaces can be found in larger government institutions like the Los
Alamos National Laboratory (LANL) Research Library (Collins 2007), the NSF Data
Interoperability Networks Crosscutting Program (INTEROP), the TARDis program at
University of Southampton (Hey 2005) or more commonly in the Social Sciences through
services such as GIS and Bioinformatics (Gold 2007). The ARL report states that
practical examples of such “virtual organization” activities may include community
workshops, web resources, and task groups.

The third critical area is policy development. The Task Force relates the recent
policy related decisions made by government agencies like NSF and NIH, and the
importance of open access to content and data sharing. This question of access can be a
tricky one to negotiate, as the science community is often itself divided by the type of
information collected; with some being federally funded and otherwise open to the
public, and other that is proprietary or collected by privately funded agencies.
Laboratories will often work with data from both sides, which can lead to contradictory
rules regarding disclosure. Business corporations may encourage some publishing of
“frontier” data by their employees in the “open access” realm, but because most corporate
research labs are ultimately profit-seeking businesses not everything will be disclosed
(David 2006). As a result, it is important that standards are universally recognized
between actors, so that legal disputes do not arise as data is shared.
The establishment of standards by government agencies, though, can only help further the release of “open access” information. As such, as ARL points out, NIH and NSF supported tools are therefore excellent early examples of how the scientific community can change licensing models to transform the scientific research cycle into a more fluid and e-Science friendly environment (Lougee 2007).

In addition to the critical areas, ARL further defined five desired outcomes, and action plans and strategies that can be utilized to reach these goals. It is important to identify these characteristics, as examination of on campus laboratories will necessarily reveal opportunities to put these strategies into practice. The outcomes identified were:

- An ongoing capacity and process within ARL to develop, coordinate, and evaluate an e-Science program agenda.

Within this outcome, ARL primarily plans to establish internal mechanisms for promoting e-Science support agendas. As this outcome is targeted at ARL itself, it is out of the scope for any sort of local structural or institutional changes. It does however, illustrate the need for a diverse body of specialists to oversee these strategies by calling for members from several committees, in addition to external stakeholders, to develop and identify policies and organizations with which to establish communication.

- A widely shared understanding within research libraries and among other stakeholders in the e-Science support community of how libraries can contribute to the development and ongoing evolution of cyberinfrastructure and e-Science.

The strategies and action plans enumerated for this outcome relate to the establishment of uniform policies that can be shared among research libraries. Similar to the above internal guidelines, these rules and policies would be crafted by both
information professionals as well as external stakeholders. It also recommends a need to ensure that library directors and university librarians understand and are committed to these policies, and can bring them to their home institutions for consideration and implementation. The action bullets recommend the use of case studies to promote e-Science programs on an institutional level, by using examples like CERN, Queensland University of Technology, Woods Hole Oceanographic Institute, and others.

- Knowledgeable and skilled research library professionals with capacity to contribute to e-Science and to shape new roles and models of service.

The strategies for this outcome are perhaps one of the most important for academic libraries, as they illustrate how ARL can provide support for the development of e-Science programs on a local level. They discuss the establishment of ARL maintained education tools such as glossaries, resource bibliographies, inventories, and wikis for use by academic librarians looking to establish programs. The priority for ARL in this example is taking the information they collect from their internal committees and sharing it with library workforces, similar in scheme to how knowledge could be shared in an e-Science grid network.

- Research libraries as active participants in the conceptualization and development of research infrastructure, including systems and services to support the processes of research and the full life cycle of research assets.

Following the establishment of programs in academic institutions, this step seeks to help sustain programs by providing an active engagement mechanism. ARL would actively monitor established programs, and collect documentation of newly established repositories. This knowledge base would not only help sustain long-term programs, but
would be useful to the establishment of new programs by smaller or newly established institutions. Specific goals also list partnership with national organizations like the Coalition for Networked Information (CNI), and the National Academy for knowledge sharing.

- Influence on policy, standards, and resource allocation decisions that support ARL principles.

This final outcome speaks to the development of the open data movement. Relating directly to the first and third critical areas, ARL seeks to address the concerns of the publication of data and to establish compliance standards for members.

Obviously, this report is of critical importance to academic libraries, especially those located on campuses with a number of science laboratories already engaged in research networks that expand beyond campus boundaries. To date there is no research showing how academic libraries are responding to the ARL Report or how well its recommendations mesh with the needs of scientists themselves. This paper examines the findings of this report, particularly noting the areas for engagement and the recommendations, and couples them with interviews conducted with science faculty at the University of North Carolina – Chapel Hill. These interviews help to establish some practical examples for engagement at an institutional level, as well as identifying additional challenges that exist in creating an infrastructure for support.
Review of Literature

Although there is not a great deal of literature available on the topic of e-Science there are several notable writers who have explored this topic area, and some case studies that can be drawn from. Further, there are related fields such as the discussion of digital repositories and ‘D-space’ that apply to the discussion at hand. In the course of researching for this paper, I tried to draw both from literature that related directly to the development of e-Science in the academic and scientific communities, as well as alternative support mechanisms such as digital repositories to develop both my interview questionnaires, as well as my recommendations for action.

The most important document currently for professionals in the Information and Library science field is the Agenda for Developing E-Science in Research Libraries from ARL. It is discussed at length in the Introduction portion of this paper, but its focus on the development of institutional infrastructure, and its recommendations for action are vitally important. Further, the Readings and References portion of the third appendix contain some valuable sources which helped me in my own research. There are two articles by Anna Gold, a member of the subcommittee that developed this report, entitled “Cyberinfrastructure, Data, and Libraries” (Gold 2007), that speak to both the processes for the development of infrastructure at an institutional level, as well as the current state of data librarianship and roles and examples that libraries can draw from in developing their own e-Science communities. Gold recommends expanding upon the Social Science research networks that already exist in most University libraries in the form of GIS data,
developing data consultancy and referral services, and participating in professional data
curation conferences to expand the body of knowledge that librarians possess so that their
professional skills will be better able to meet the evolving needs of research scientists in
an academic community.

Another author recommended by the ARL report is Tony Hey who has written,
among other related materials, articles entitled “e-Science and its implications for the
library community,” and “Cyberinfrastructure for e-Science” (Hey 2006 and Hey 2005).
The first article explores the basic issues surrounding the growth of e-Science, and what
steps need to be taken by the library community to support its growth. Drawing upon
case studies such as the CombeChem, eBank, and SmartTea projects, Hey points to
models like the TARDis project at the University of Southampton as an example of how
institutional repositories can expand the role for librarians to participate actively in an e-
Science community. The essence of the TARDis project is the establishment of a
repository that serves as both a traditional repository as well as an open access archive.
Hey writes that, “University libraries… may need to take responsibility for assisting with
the curation and preservation of smaller scale data sets arising from the research of
research groups or individual academics,” (Hey 2006). This sentiment is echoed later by
one of the faculty involved with this project. The second Hey article,
“Cyberinfrastructure for e-Science” looks at distributed computing infrastructure for
advanced science and engineering, also known as Grid technology. Web service grids
are used to establish virtual organizations similar to those discussed in the ARL agenda.
Software clients make it possible for researchers to communicate and share data across a
unified interface. This article discusses several projects that use Grid software, particularly in the UK.

Several other authors have written supporting papers about the challenges that information professionals will be confronted with as they begin to expand the role of the academic library to support e-Science networking. The National Consultation on Access to Scientific Research Data (NCASRD) has issued reports about both the challenges to open access, particularly concerning legal constraints surrounding private data (Strong 2005). The concept of “open science” is expanded upon in an article by Paul David, Matthijs den Besten, and Ralph Schroeder entitled “How open is e-Science?” (David 2006). This article further discusses the challenges behind establishing the “openness of data”. Ian Foster has written an article entitled, “Service-Oriented Science” which explores how tools can be best designed to enable the widespread automation of data analysis and computation (Foster 2005). By providing descriptive models, it discusses how software development can support growth, and provides examples of some academic programs like Purdue Universities NanoHub and UCLA’s Grid are becoming campus versions of discipline-specific or national programs like the UK’s Grid Project (Hey 2005), or the US Dept. of Energy’s Earth System Grid (Bernholdt 2005) and Fusion Collaboratory (Keahey 2002) projects.

These projects can be paired with developments and explorations into the establishment of digital repositories through use of software platforms such as Dspace. Dspace is freely available open source software that can be used to capture multiple formats of data, distribute this data over the web, index material so that it is searchable by other users, and preserving the data over time (www.dspace.org). There have been
several case studies done of libraries that are using Dspace on varying scales, and Dspace could easily be used as another example of a virtual organization that could be established by an academic library to support the curation of scientific data. Case studies include a pilot project done by the Ontario government (Devakos 2008), the use of Dspace by several Indian universities (Vijayakumar 2006), its use by Texas A&M (Weimar 2006). Dspace is also often mentioned in papers discussing the establishment of institutional repositories (Hixson 2007).

Dspace is just one tool that can be implemented to establish a virtual organization, and thus should be considered one of many options available. There are articles that explore reasons why Dspace has not always been a successful tool in promoting such a community in places like Cornell University (Davis 2007). Ultimately, an institution will need to decide which tools are the best fit for their specific needs.

Due to the recent release of this ARL report, and the relatively small number of articles available on this topic, this paper will attempt to illustrate the specific needs and observations that scientists have on one campus. In order to begin the process of establishing an on campus e-Science community, it is important to carefully examine the many individual needs that the variety of research laboratories will encompass.
Methodology

This paper looks at the impact of e-Science at one particular university, the University of North Carolina at Chapel Hill (UNC-CH). According to the information page (available at http://www.unc.edu/news/compendium.shtml), UNC-CH currently enrolls more than 27,000 students in 71 bachelor’s, 107 master’s, 74 doctorate and four professional degree programs. The University currently has 3,200 tenured faculty members, and consistently ranks among the top 10 public universities in *US News and World Report*. The University library system circulates more than 5.8 million books and holds over 54,000 serial subscriptions in a variety of disciplines. In 2006, the University faculty was awarded nearly $600 million dollars in contract and grant funding, with a significant portion coming from the National Institute of Health (NIH) to support UNC medical research programs, and another large portion being devoted to genome sciences and cancer research.

This paper is an examination of the previously discussed ARL report “Agenda for Developing E-Science in Research Libraries”, which addresses key areas for concern among academic librarians, and strategies for improving support on institutional levels. For the purposes of this paper, I will be carefully considering the critical areas for engagement, and the second, third and fourth outcomes. These outcomes in particular relate to how academic libraries can establish programs to provide support for the e-Sciences (the first and fifth are primarily directed at institutional changes within the ARL).
The ARL report is intended to address the needs of all research libraries, but I was interested in how applicable the report might be to individual scientists at one university. I chose to conduct brief interviews with several scientists from research laboratories on the UNC-CH campus to establish what some of their individual needs would be should the library move to a support role in their individual e-Science networks. Since this is exploratory research on a topic that has not been covered before, I chose to gather the data by interviews in order to gather richer data. In addition, since I wanted to include a variety of on-campus laboratories I thought it would be difficult to establish a common set of questions that would be universally applicable.

I developed an interview questionnaire that can be divided into five distinct sections. Each section relates to a different aspect of e-Science partnerships: Affiliations (which other actors this individual lab includes in its research networks), Data Collection (what processes each lab follows for data collection, and what space considerations they have), Data Distribution (how their data is handled and published following collection), Security (what security considerations need to be made while handling this data), and Library Association (what preexisting relationships exist between this lab and the University library system at UNC-CH). The interview questionnaire is located in Appendix A.

The questions in the survey were developed to gather information relevant to the ARL report and each portion of the interview relates to different sections of the ARL agenda. The first, affiliations, seeks to establish the individual networks in which the laboratories participate. If the University library is to provide support for these laboratories it will need to be aware of the size and scope of these networks, as they can
encompass a large number of external sources. In these networks lie opportunities for collaboration with not only the individual lab, but with university libraries from other institutions, government agencies and organizations, and other departments on campus.

The second group of questions relate to data collection; the questions there speak to the individual research policies of the laboratory, and the amount of data they collect individually. If the library plans to serve as repository, it must be capable of storing a portion of the data collected in some manner. Further, if the library wants to establish standards for data storage it will need to know about the sources from which the data is gathered. Issues of proprietary rights and public vs. private domains mentioned in the policy development area for engagement are addressed in this section and the subsequent sections on distribution and security.

The section relating to data distribution addresses what measures each lab already undertakes to publish its data, either through publication in journals or on websites. It also addresses the frequency with which information is published or otherwise made available, and which parties within the lab are responsible for making it available. The questions about security address the potential confidentiality issues that may arise with collected data, and the current security measures each lab employs to protect its data prior to publication. The last section on library association addresses what current relationships exist between the individual laboratory and the UNC-CH library. This section addresses what, if any, existing library resources the labs regularly utilize, and what opportunities they feel there may be for collaboration between their lab and the library system.
Once the questionnaire was developed and approved by the IRB, I began to develop a contact list of research labs on campus specifically within the sciences. One of the early difficulties I ran into was that there was no dedicated listserv, due both to the multi-departmental distribution of research labs, and due to the sheer number of laboratories. I utilized existing web links from departmental home pages, and accumulated contact information from specialized librarians to create a list that contained labs from Biology, Chemistry, Geology, Marine Sciences, and Biomedical Engineering. Once the list was completed I sent out emails requesting interview time either in person, over the phone, or via email correspondence.

After a substantial period of little response, I was able to establish contact with five University laboratories from a variety of disciplines. The respondents will be referred to hence forth as Lab A, Lab B, Lab C, Lab D, and Lab E. I will briefly describe each lab for context:

Lab A is a research laboratory in the Biology department that was founded in 2001 when its coordinator joined the UNC faculty. This lab specializes in genome evolution and the genetics of complex traits, with a focus on plant life. The staff includes both Graduate students at UNC-CH, post-doctorate assistants, and full time staff members that help to maintain the lab’s two websites. Of these websites, one is continuously updated (on a daily basis) with new information, while the other is updated annually with research findings. In the past, this lab imported most of its data from external sources, but has recently begun to produce a substantial amount of in-house data sets. Data compiled by this lab includes genomic data and sequences, gene expression
data, genetic variations, and phenotypic measurements. The interviewed representative from Lab A is the director for this lab.

Lab B is a research laboratory, founded in 2000, located in the Marine Sciences department. This laboratory deploys remote sensors off the coast of North Carolina to collect data, in coordination with several other national and academic laboratories that collect similar suites of data. The work this laboratory does is divided into three distinct areas: data collection, which involves delivering remotely collected data to the shore for analysis; modeling, where this data is used to create numerical forecasting models; and data management, where this data is aggregated and reported in real-time to other agencies. The data that this lab collects is primarily physical oceanographic data, which can include measurements of currents, water levels, salinity (the concentration of salt in water), temperature, and waves. This laboratory generates a significant amount of data when combined with the raw data from other actors, and uses three websites to publish its findings. The interviewed representative from Lab B is the director for this lab.

Lab C is affiliated with the Institute for the Environment (IE) and the Geological Sciences department. It was created through a private donation to the IE in 2007, and opened in early 2008. This laboratory currently works with several government organizations (US Department of Energy, NSF, EPA, NOAA), and for now one other academic institution, though it hopes to soon begin collaborating with more universities in the near future. This lab collects experimental (field) data from climate related sensors, seismic data from Arctic field work, and computer output from climate simulations. This lab currently does not have its own website, but is hoping to produce one in the near future with University IT support. Currently the director of this lab
manages all of the collected data, and the findings (data visualizations) of this lab are published using science journals. The interviewed representative from Lab C is the director.

Lab D is a smaller laboratory from the Chemistry department. It was created in 2004 when the director joined the University, and is currently staffed primarily by Graduate students from the Chemistry department. Its research focus is on the design and synthesis of polymers and the design and fabrication of polymer composites. As opposed to the other laboratories, it does not produce a large amount of data to store, instead producing physical compounds, and measurements. The research from this lab is reported through either serial publications or conferences. It does not currently have a website, aside from a departmental information page. The interviewed representative from this lab is a graduate student who was recommended by the director.

Lab E is a similarly small laboratory from the Chemistry department. This lab specializes in Organometallic Synthesis analysis – Nuclear Magnetic Resonance (NMR) data, IR, UV-Vis, Crystal diffraction, Mass Spec, and elemental analysis. The Organometallic Synthetic Lab was started in 1976, and predominantly employs graduate students, with occasional undergraduate or post-doctoral assistants. This lab also does not have a specialized website, and utilizes a departmental link to a basic info page. The data that is collected by this lab is published in journals, frequently as embedded images, and not on a particular webpage. The interviewed representative from this lab is a graduate student who was recommended by the director.
Analysis

Despite the different departmental affiliations of each of these labs, there were several similarities that were found in each of the sections. Comparing the data side by side, there are several trends that can be established. However, due to the small number of interviews, the findings of this paper do not necessarily represent University wide trends, but are more appropriately considered as examples from which further investigation can be done.

In the first section on affiliations, it becomes fairly obvious that there is a great difference in the number of external affiliations on a case by case basis. Lab A (stating too many University relationships to list), and Lab B (six other Universities that they correspond with on a daily basis) were both involved with a larger number of other Universities and government agencies, while the Labs C, D and E have fewer regular contacts with other academic or governmental agencies. Within the University, departments were generally connected through Graduate students, as opposed to a great deal of cross-discipline focus in research. In both and Labs A and B, it was mentioned that Graduate students often lend expertise from their home department, but did not reflect a focus in study. For example, in Lab B a graduate student from the Geography department opened up opportunities for collaboration through his personal contacts and subject expertise.

The relationships between other Universities and government agencies reflect some of the challenges that the ARL list in the section on policy development. Because
each lab interacts with entirely different organizations, it will be necessary for the University library to establish standards that will be acceptable for use among organizations. Further, academic standards that vary between disciplines will need to be considered as each will likely have different rules for both the organization of information as well as government standards that may be established by national organizations.

In the data collection section, each lab receives data from both in-house sources as well as external sources to some degree. Though the Lab D and E labs received the least amount of external data (Lab D primarily receiving toxicity reports on compounds and Lab E sending information to and from a Microlab in Kansas) all five receive data from both academic and private sources. As a result, though some of this data is public or potentially “open access” data, some is still purchased or proprietary data. Again, this will vary from lab to lab, but highlights the need for flexibility in library policy as no two labs will be exactly the same. Differences in data production will have the most impact in the areas of Data Issues and Policy Development by the ARL Agenda. From a repository standpoint, the University library system will need to be able to differentiate between the sources of data, while at the same time maintaining general standards.

In terms of storage, Lab D was quite different from the other three, in that their production was generally not measured in stored data, but in physical compounds. As such, they use very little storage space, as opposed to the Lab A, B and C which can generate close to a Terabyte of data per year. Lab E, another Chemistry-based lab, collects data for publication that is stored on UNC server space, but was unable to speculate as to the amount.
Lab B generates even larger amounts of stored data, as it collects information from several sources concurrently. The sensors alone only produce .5 gigabytes per year, but the numerical models they generate roughly 7 terabytes of data per year in storage, and accumulated data from networked institutions that utilize satellites generate as much as 100 terabytes in a year. This sheer amount of incoming data would pose a significant obstacle to long term storage needs, although it would open up different opportunities for support that will be illustrated later. For the research labs that use significant amounts of space for data storage, each uses a combination of on campus server clusters provided by ITS, as well as a variety of off-campus storage methods. For some, like Lab A this includes storage by other institutions or facilities provided by grant money. This again is a concern should the University wish to serve in a curatorial role for this data, as it would reside in several locations.

There was a stark difference between labs in terms of distribution methods, with Labs A and B making extensive use of their websites to publish their data sets. The C and D Labs both have websites in development, though Lab D uses regular presentations and serial publications to publish its data. Publication becomes a serious consideration while addressing the virtual organizations ARL seeks to establish. Data in Labs A and B lie predominantly in the public realm, and can be published quickly on their websites. Information gathered in the Labs D and E, however, is proprietary and is kept private until it is ready for publication. Lab C demonstrates some of the difficulties present in its data collection, as some is mandated by the NSF to be published within a certain time frame to the public domain, while other data is private and cannot be shared. When speaking with the coordinator of Lab B, he echoed a similar sentiment. While the
majority of the data his lab collects belong in the public realm, there are occasionally pressures from private institutions to keep portions of data out of the public realm.

The Agenda established by the ARL places a great deal of importance on the concept of “open access” and virtual organizations that researchers can use to share data in real time. Librarians are accustomed to working within the restrictions of copyright. In a similar manner, they will need to learn to establish standards for proprietary rights to data as networks that research labs belong to will generally receive data from a variety of sources.

None of the scientists I interviewed viewed security issues as a great concern, as none collected confidential or sensitive data. Each lab had its own variation of password protected documents and databases to keep data from being manipulated, but none were responsible for data that could be considered confidential or identifying (data such as medical records, social security numbers, or other identifying information). This will be a larger concern with Biomedical Labs, and other research labs that use human subjects as part of their data collection process.

The final section of the interview that asked the lab representative about the University library provided an indicator of a uniform trend. Namely, none of the research labs maintained any form of regular contact with the University Library system, nor do they use the available online resources outside of e-journal subscriptions. The Lab D representative mentioned frequent use of the Chemistry Library facilities, but again this related to journal use, and not to a regular interaction with a librarian. The Lab E representative wrote that the majority of their research work was done through use of e-journals, and expressed a preference to electronic serials as opposed to print. None of the
labs regularly share their information with the Library system, none maintain active linkage between their own sites (for those that do have websites) and library websites, and none mentioned contacting the library for research support (purchasing journals or databases). The coordinator from Lab A mentioned subscription to two databases, but was unsure from where the specific funding for access came.

From a librarian’s point of view the results of the final section that shows the current absence of interaction existing between the UNC-CH library system and individual research labs is disappointing. Although the small number of interviewed representatives cannot indicate a university-wide trend, it shows that at least these five research labs conduct their day to day operations beyond the library’s scope. They obtain their information from in-house experimentation and data collection, and from external consortia shared repositories. They interact with national and academic agencies across the country in order to compile the data they require. However, they do not utilize library resources outside of the most basic interactions of using physical or electronic materials to supplement research.

Obviously, there is a need to foster stronger ties between the labs and the libraries. As a final question, I asked each representative where they felt there were opportunities for collaboration between their individual labs and the UNC-CH library system. These responses were interesting because they often mirrored some of the goals for the ARL Agenda.

The representative from the Lab D felt that the lab was fairly self-sufficient as is, and did not have a great deal of need for additional library support. He specifically mentioned missing the convenience of a departmentally situated specialized collection,
however (the Chemistry library having been moved to a combined location with the Biology, Botany, and Zoology libraries). He did show interest in the concept of specialized research guides, a resource he was unaware the library system currently offers. These research guides can be found at http://www.lib.unc.edu/guides/ and are divided first by subject area, and further into tutorials by subtopic. As Lab D currently does not have a website outside of a basic informational departmental website, it does not maintain linkage to the University library website, but as this website is developed it may present an opportunity for collaboration.

A similar opportunity exists with Lab C, as its director specifically listed the development of a website as an area in which he would welcome library support. Lab C is currently developing its website, and library support could be as simple as maintaining linkage, or as complex as developing research guides to support the data sets the Lab C will publish.

Lab E recommended the subscription to more electronic journals, but in an earlier section note that their website is currently not updated. As it appears as though the lab technicians do a majority of their research work remotely (based on their responses to the final two questions), perhaps an opportunity exists in developing more web based support for a specialized laboratory.

The coordinator from Lab A suggested a support mechanism that was not specifically tied to research lab support, but more to the support of individual university faculty members. His recommendation was that the university library system establish a digital repository to help faculty members curate their personal data sets. Faculty members generate their own data collections, but often lack the resources necessary to
properly store and maintain them. Stating that faculty produced material often dies due to lack of curatorial support, he felt there is a great opportunity available for libraries that already utilize institutional repositories for printed materials to similarly care for digital materials. He further noted that making domain specific librarians available for support would generate more interest from departmental faculty.

Lab A’s coordinator presents an interesting opportunity to develop virtual organizations within a University in a way that transcends individual laboratory research. An institutional digital repository for faculty would be a resource that could publish “open access” information, make it searchable and organized outside of the university, and assist the knowledge discovery process by producing materials that might otherwise be lost. There would need to be substantial financial support from the University library system to create such a repository, and then on the part of individual faculty to submit materials for publication.

Lab B’s coordinator presented a different sort of opportunity that related predominantly to special skills that exist within the Information and Library science profession. In his lab, a great deal of raw observational data is constantly collected (both in house, and through external sources). Within Marine and Earth Sciences, there are constantly evolving standards and ontologies created to make this data searchable. The great fear, though, is that remote labs are constantly in danger of replicating efforts that are being undertaken in other parts of the country to develop these standards. Within the Information and Library science profession, there is a high level of expertise regarding the creation of ontologies and existing metadata standards. This expertise could be lent
on both local as well as national levels. This speaks directly to the policy development area ARL has identified as a critical area.
Conclusions

The information gathered in this paper is a starting point for further research into how a university library system can provide support for on campus research laboratories. Keeping in mind the ARL guidelines, the following areas can be identified through the case studies in this paper:

Establishing a committee for e-Science collaboration

The first and most obvious step for the development of collaboration within the realm of e-Science is for the library system to establish its own infrastructure for promoting collaboration. Following the model provided by the ARL, such a committee should include librarians, administrators, and faculty members from within the individual science disciplines. As each will have unique perspectives on the issues facing these research networks, it will be important to have all sides represented in the founding body.

This committee would be responsible for establishing university wide standards for data sharing, improving outreach towards the scientific community, developing relationships with similar bodies outside of this institution, and creating/making/purchasing resources to promote a university research infrastructure.

With such a body established, it would be much easier to initiate new programs such as a digital repository for faculty, or promoting new data standards for “open access” faculty data sets on an institutional level.

Website development and support
Two of the scientists interviewed for this project stated that their labs were in the process of developing websites, while two maintain several sites that do not currently share links with the University library system, and another has a website that is currently not regularly updated. Assisting labs in the creation of content for their individual websites is a minor effort that can be undertaken by the library with support from ITS. By simply sharing links with the University library homepages these labs are expanding their networks to include the library system. This linkage would assure that their specialized resources (for example, the resources and data currently provided by both Labs A and B) could be found while exploring departmental library websites, giving more students exposure to their data sets. As graduate student populations create collaborative opportunities between departments, as demonstrated by Lab B and their relationship with the Geography department, opportunities for such relationships do not always exist.

Further, with the current development of specialized course pages, tutorials, and research guides, there may be opportunities for the library system to publish materials that can assist the research process of an individual lab (by giving it a one-stop location for finding related library resources), or by creating materials that can supplement specialized materials for students of different backgrounds.

**Development of shared contact lists and liaisons**

During the course of my research I had a great deal of difficulty locating an existing resource for contacting faculty members who maintain research labs on campus. The representative from the Lab D recommended using a departmental listserv to contact lab technicians from a particular discipline, but a combined listserv for research labs
would provide the library system with a mechanism for outreach. Such a listserv could be utilized to advertise products or services, to solicit opinions or questions from research labs, and to generally create a network that includes the University library system.

It seems as though a significant reason for the lack of collaboration between the library system and research labs derives from a general lack of communication. Most labs will not contact their departmental libraries directly, and do not utilize resources except for the most basic services. As such, there is little need for them to share new information or data with the library system. Avenues for communication could be established simply by having the preexisting specialized departmental librarians/liaisons create contact lists for research labs within their departments, then sharing these lists system wide. Establishing regular contact with the research labs may also be an opportunity to develop individual relationships that can serve as bases for further support. Even if faculty members do not participate immediately, at least a system for communication will have been established for future contact.

**Developing repositories for faculty data sets**

Following the recommendation of Lab A’s coordinator, there is an opportunity available for the establishment of an institutional repository for digital data sets that are generated by faculty members. There would be significant costs associated with such a project, but a more intensive study to gauge interest in such a project would provide the library system with a progressive opportunity to provide support to the faculty in a manner that would mirror ARL initiatives on a national level. By examining models such as the TARDis project (Hey 2006), UNC-CH may be able to find a means to establish a
similar repository/open access forum for faculty to publish materials or data that may otherwise never be published.

Lending expertise

Following the coordinator from Lab B’s suggestion, there are currently opportunities for librarians to lend their expertise, particularly in the realms of data management, and the establishment of standards. The difficulty in this recommendation is that such efforts would need to be done on a case-by-case basis, and would most likely require an individual lab to present the opportunity. This furthers the recommendation for the establishment of regular contact between the library system and research labs on campus, as opportunities like these will become evident through increased contact.

The most progressive step that the library system at UNC-CH can take is to establish a forum for communication with individual laboratories. Whether this is through a unified website or similar web resource, or through simple measures such as the creation of a dedicated listserv or departmental mailings, it will be the first step in a long process. By establishing a ‘virtual organization’ in addition to administrative infrastructure for the development of e-Science on campus, UNC-CH will have a unique opportunity to create a new style of research environment that transcends traditional campus boundaries.

Suggestions for future research

As stated, there is still a great deal of data collection to be done for any sort of e-Science initiative to be undertaken at UNC-CH. This paper is exploratory by nature, but shows that there are several specific areas in which future research can be done to support the underlying goal of creating virtual organizations on campus. A comprehensive
survey distributed to laboratory directors and their employees asking for descriptive data about their individual environments could help to establish the current shape of the research laboratory environment on campus. How many labs are there in total, how many employees/graduate students/post doctorate assistants are employed at each, how many currently maintain websites/databases, how can these labs be contacted, etc. A survey of this sort could be very beneficial to library administrators as they plan for future infrastructure to support an e-Science community.

Surveys or interviews could also be conducted with universities like Cornell, Purdue, and UCLA that have already taken steps towards creating an interactive virtual organization on campus through use of digital repositories. Looking at these universities as case studies, and determining what steps would be most effective for a campus like UNC-CH would also provide a great benefit to campus administrators.

This paper has been a first attempt to gather data for the UNC-CH campus, and hopefully illustrates both the opportunities as well as the obstacles that currently exist towards developing e-Science support on campus. Such development will be crucial, as laboratories will only continue to develop more interactive relationships with other scientific communities across the world. More research on this topic will be required in order to make informed decisions about what measures to take towards developing institutional infrastructure to support e-Science initiatives like digital repositories or other virtual organizations on campus. Librarians need to keep up with the changes in both technology and communication, and as time goes on it may become more difficult to develop cooperative relationships with a community that may not feel that it requires
library support. The university library system must make the first step, and be proactive about developing these relationships, or risk being left behind.
Bibliography


Appendix A

Chris Granatino
IRB Proposal

Interview Questions

**Affiliations**
Name of Participant:

Lab Affiliation:

Departmental Affiliation:

Which UNC departments are represented in your lab, either through membership, or data contribution? (By data contribution, I refer to data sets that are collected by other departments and sent to this laboratory for use.)

Which, if any, other Universities are represented in your lab, either through membership, or data contribution? (By data contribution, I refer to data sets that are collected by other departments and sent to this laboratory for use.)

Which, if any, other participants (Government sponsored or private institutions) are represented in your lab, either through membership, or data contribution? (By data contribution, I refer to data sets that are collected by other departments and sent to this laboratory for use.)

Please briefly describe the history of your lab.

**Data Collection**

What kind of data does this laboratory collect? (If this question is confidential, specifics are not necessary. A response of 'confidential' will be suitable.)

Do you collect your own data, or do you receive collected data from external sources?

Is data collected within a consortium, or obtained through other means (purchased, etc)? If so, are the actors in this consortium listed above? How is this consortium structured?

Who has ownership over the collected data? Is this data used solely by this lab, or by others? (Possible follow-up) How is ownership negotiated through a consortium?
How much (if possible to estimate) data is collected annually (in gigabytes of data)? Monthly?

Is data stored locally (on UNC Server space, or on UNC computers), or externally (off-site Server or storage facility)?

How much data is currently stored (if available, and not confidential) in gigabytes?

How is data stored (Servers, server clusters, hard-drives, etc.)? What are the costs associated with the storage of this data? (If available)

**Data Distribution**

How is this data distributed, if at all? (Transmitted to external sources, published for private use, published for public use?)

How frequently is this data updated/changed/deleted?

If networked, does this lab have the ability to alter data that other labs collect? Can other labs alter data that this lab collects?

Who makes updates/changes/deletions?

Does your lab have a website?
URL:

Are your data sets available through this website?

Are they available online, either free to the public or through a paid/subscription based service?

Who maintains the website? Follow-up: (If the lab makes their data sets available on a secondary site) who maintains the site where the data sets are made available?

**Security**

Is the data collected by this lab confidential, or does it contain sensitive data?

Is data altered after collection to protect subjects? If so, who alters this data?

What security measures does this lab currently employ to protect this data? (Specifics not necessary, an answer of confidential is suitable).

At what point is collected data sharable, if ever?

How long is data stored? Is it then made public or destroyed?
Library Association

Does this lab currently link to/from the University Libraries at UNC?

Does the lab currently publish guides to using their data sets? Do they link to other University Library produced guides?

Does this lab have a regular contact with a Librarian from the University Library system?

Does this lab currently share current or archived information with the Library system?

Does this lab currently utilize Library resources (purchased databases, materials, human resources, space) to conduct its research? (Follow-up) If this library is a member of a consortium, do other members have established levels of contact with their libraries to support their research?

Are there any specific ways in which you feel the University Library system could provide better support for your lab, or for research labs in general?