IN THIS ISSUE

The theme for this issue of Synergy: News from ARL Diversity Programs is “research” and how libraries are supporting these endeavors in higher education and becoming invaluable partners in the enterprise.

2000–2001 Leadership and Career Development (LCDP) Fellow, Angela Lee (University of Washington), discusses current trends in data curation and management, specifically in the context of health sciences librarianship at her institution.

Also in the health sciences arena, Myrna Morales, 2009–2011 Diversity Scholar, defines Community Based Participatory Research and the role that the research librarian can play in this new paradigm—from support for or engaging in the methodology—to providing preservation services for the data collected.

Last, Minglu Wang (Career Enhancement Fellow, 2009) writes about her experience as the newly appointed data services librarian at the John Cotton Dana Library of Rutgers University. Wang makes the case for library and information professionals insinuating themselves into research projects much earlier in the process in order to help researchers develop more efficient and comprehensive data collection and management plans.

All three of these authors provide solid evidence of the value that library and information science professionals bring to their organizations, and of the significant contributions that former ARL Diversity Programs participants are making to the profession.

I hope you enjoy reading about the important work being done in libraries by former ARL Diversity Programs participants.

For more information about ARL Diversity Programs, please visit:

www.arl.org/diversity/

Mark A. Puente
Director of Diversity Programs
Association of Research Libraries
mpuente@arl.org
About the ARL Diversity Programs

The **ARL Initiative to Recruit a Diverse Workforce**, funded by the Institute of Museum and Library Services and ARL member libraries, offers a stipend of up to $10,000 to each ARL Diversity Scholar in support of graduate library and information science education. ARL Diversity Scholars participate in the annual ARL Leadership Symposium, a research library visit hosted by the Purdue University Libraries, and a mentoring relationship with a research library professional. For more information about the Initiative to Recruit a Diverse Workforce, visit [www.arl.org/diversity/init/](http://www.arl.org/diversity/init/).

The **ARL/Music Library Association (MLA) Diversity and Inclusion Initiative**, funded by the Institute of Museum and Library Services (IMLS), MLA, and ARL member libraries, seeks to recruit diverse students with advanced degrees in music to careers in academic and research libraries. The program offers tuition support and a paid internship for up to one year in one of five partner music/research libraries. Other components of the ARL/MLA DII include a formal mentor program and support to attend the MLA annual conference where participants will receive specialized instruction in the areas of career development and effective job search strategies. More information about the program is forthcoming, but an announcement concerning the IMLS award can be found at: [www.arl.org/news/pr/ARL-MLA-DII22june11.shtml](http://www.arl.org/news/pr/ARL-MLA-DII22june11.shtml).

The **ARL Career Enhancement Program (CEP)**, funded by the Institute of Museum and Library Services (IMLS) and ARL member libraries, offers MLIS graduate students from underrepresented groups an opportunity to jump-start their careers in research libraries by providing a robust internship experience in an ARL member library. Each CEP fellow participates in a six- to twelve-week paid internship in an ARL library, a mentoring relationship with a professional librarian while on campus for the internship, and an opportunity to attend the annual ARL Leadership Symposium during the American Library Association (ALA) Midwinter Meeting. There is also potential for academic credit for the practical intern. For more information about the ARL Career Enhancement Program, visit [www.arl.org/diversity/cep/](http://www.arl.org/diversity/cep/).

The **ARL Leadership and Career Development Program (LCDP)** is an 18-month program to prepare midcareer librarians from underrepresented racial and ethnic groups to take on increasingly demanding leadership roles in ARL libraries. The LCDP design includes: two LCDP Institutes, an opening and closing event held in conjunction with national professional meetings, a career-coaching relationship with an ARL library director or staff member, and a personalized visit to an ARL member library. For more information about the LCDP, visit [www.arl.org/diversity/lcdp/](http://www.arl.org/diversity/lcdp/).

---

**Synergy: News from ARL Diversity Programs** was first published in 2007 as a forum in which to share information about programs and opportunities centered on careers in research libraries.
Data Stewardship from a Health Sciences Library Perspective

ARL has been encouraging libraries to take on the challenge of data management since tracking the trend in eScience in the last decade. As expected, data curation and data stewardship are now current buzzwords in the library community. Along with journals, data has become the unit of information currency. For academic research, they are the primary source of new knowledge. It’s hardly surprising given that science in the 21st century has been transformed by our technical ability to do high-performance computing on massive sets of data. The library literature is replete with reports and studies on the subject but getting a handle on the main concepts and issues can be daunting to initiates.

The purpose of this article is to provide an overview of data management with special emphasis on health research. It focuses on what we as librarians should know about data management and curation, what role we can play, and how we can support researchers. At the University of Washington Libraries (UWL), we have put together a Data Services Team (for which I serve as the Health Sciences Libraries representative) to begin developing and implementing a library data service program. While the UW certainly provides data services to its researchers (e.g., GIS, social science data, bioinformatics) they are offered by different units on campus (UW Libraries, Center for Social Science Computation & Research, and eScience Institute). A first goal for our library data services program is to build, store, and share data and to act as an intermediary for other data providers on campus. It is a whole new role for librarians, and, as always, the very first step to success is knowing what we are getting ourselves into.

Some Fundamental Concepts of Data Management

Data Life Cycle (DLC): Understanding the DLC is key to knowing where libraries fit in the management of data. The DLC is the prequel to the publication life cycle. It focuses particularly on the pre-publication stage of the research process, starting from data collection to findings. A typical research process flow goes something like this: Hypothesis > Experiment > Analysis > Results. However, that process can vary greatly. Not all health researchers, for instance, conduct pure experiments. In some cases they base their hypotheses and findings on information they mine from clinical databases. The challenge is to determine where librarians appropriately fit into that research process and how we can help researchers at different stages. Librarians rarely operate in these early phases of research, nor are they required to address issues associated with the DLC. In the case of health research, librarians must also be able to deal with issues such as privacy (HIPAA).

Data literacy: A basic understanding of data in all of their forms and uses is critical for data management. Data literacy involves knowing how to access, view, manipulate, analyze and interpret data. Data literacy is an integral component of information literacy. It will be a significant challenge for libraries to build a skilled, data literate workforce. While librarians are not expected to be experts or to teach users quantitative methods, they would need to acquire at least a basic understanding of statistics, methodology, and data applications to be able to instruct users on how to locate and evaluate data sets.

Domain knowledge: Domain knowledge is not a prerequisite for data management, but it is helpful to understand how data collection is structured and preserved in different disciplines since every field has its own kind of data and data handling practices. The sources of health data are diverse, but they can be broken down into four main types: medical records, certificates of health (vital statistics), health surveys (and surveillance), and research studies (clinical trials, genomics). Data in medical records can be further subdivided into personal (name, age), clinical (medical history), administrative (referrals), economic (insurance), and behavioral (medications adherence). By being aware of the kinds of data available, we get to know the data requirements and
best management practices (i.e., what to collect, where to store, how to access, when to maintain) for that academic field.

Data management plans/guides: A data management plan (DMP) is a document that outlines how a researcher plans to manage data during and after a research project, including how data will be organized, maintained, and shared (http://guides.lib.washington.edu/content.php?pid=145215&sid=1407802). Many funding agencies including NIH and NSF require researchers to submit a formal DMP when applying for grants. A number of DMP guides are available for researchers in drafting a plan. But ARL also recognizes this as an opportunity for librarians to play a role in assisting researchers. A full discussion on this topic is available on the ARL website: http://www.arl.org/rtl/eresearch/escien/nsf/.

Data archiving: The institutional repository (IR) is where libraries can make their greatest contribution in data management. Libraries can offer scholars long-term, centralized storage of their output and manage large amounts of data that currently reside in their local departments or institutes. Libraries can also provide efficient means of preservation, metadata creation, and tools to access data. At Purdue University Libraries, the Data Curation Profile Toolkit is specifically designed to assist librarians in starting their own data interview process. In health research, where the focus is on clinical research, a comparable repository is the clinical data repository (CDR). Typically, all clinical institutions have a CDR that might host medical records, laboratory results, diagnostic reports, etc. At the UWL, the IR exists as part of our Digital Initiatives Program, ResearchWorks (http://digital.lib.washington.edu/), which is made up of three distinct services: Archives, Digital Collections, and Journal Hosting. The IR uses DSpace, an open source software from MIT, that collects scholarly output in different formats. Researchers can now directly deposit their data sets and textual reports into the Archive.

Data collection: Librarians can play a huge role in the selection, acquisition, and licensing of data and data sets. Data is a hot commodity that can be linked, manipulated, used and reused for multiple purposes, but, often, at a price. Librarians must be prepared to allocate funds in their budget and to negotiate licenses for data collections. In addition, there are access, storage and maintenance issues to consider. As publishers begin linking data to journal publications and as departments learn of their availability, libraries will have to come up with an adequate cost model for the purchase of new collections. At the UWL, attempts are being made to build its data collections but the purchase of data sets are ad hoc at best. Given the “newness” of data collections, there is no official budget line for them although this is expected to change in time. There are, of course, a lot of data that are freely available. In health research, for instance, researchers can access free government data in the form of health surveys. However, clinical data are closely guarded by clinics and hospitals, and health industry data are proprietary.

Data sharing: A number of questions come to mind: Why should we share data? If we share data, how can we make them available? Through a repository, journal, website, peer-to-peer? How will people find the data? What tools (applications, format) are needed to access the data? Are there any limits or restrictions on data usage? Are there ethical concerns or privacy issues? How should data be cited? All these issues are under discussion by the UWL Data Services Team. Our IR has developed some guidelines to assist researchers in making those decisions: http://digital.lib.washington.edu/faq.html. One of the primary concerns about sharing health data is getting access to sensitive data. Whenever patients or subjects are involved, issues on privacy and confidentiality are paramount. This is where Human Subject (IRB) enters in. Strict review, protocol, and legal measures must be in place before one can get access.

Data providers must ensure that sensitive data are securely stored and safeguarded. A couple of strategies for making data more accessible are to anonymize the data or to place data in safe harbors where they can be accessed via secured networks.

In summary, 21st century scientific research is data-driven. To ensure libraries’ continued relevance, we will have to adapt to the changing academic landscape and to take on new tasks and new roles to support the current needs of our researchers. Acquiring the knowledge and skills for data management are the newest challenges for libraries and librarians. Many other issues relating to data management, such as data security, data linkage (interoperability), data ownership (IP), and data standards could not be covered in this brief article. To learn about these topics and more, I recommend that you read the articles listed below as well as explore the emerging work at Cornell, Purdue, and Johns Hopkins.

Further Reading:

Community Based Participatory Research

On November 20, 2010 I attended a film screening of The Deadliest Disease in America, a film about health disparities in the United States. This film was then followed by comprehensive workshops to highlight health inequities in our health care system. One of the workshops I attended was called Participatory Health Research. Dr. Gia Barboza, a presenter and a professor of African American studies and health sciences at Northeastern University, demonstrated the need for data transparency in health research. She examined claims by the State of Massachusett that achievement gaps between those of different races have narrowed in recent years. Through application of statistical analysis, she did find that within each racial/ethnic group strides had been made, however the data presented provided evidence that not only has there been no improvement between racial/ethnic groups, the achievement gap, in fact, has increased.

During that session, Milagro Grullón of the mayor’s health task force of Lawrence, MA echoed Dr. Barboza’s remarks and made her case for academic researchers to involve members of the community in the research process. She spoke of this particular case study: “In 2006, another national research group came to Lawrence, again having developed a research plan without consulting the community. The intent this time was to measure the health effects of high levels of pollution. The researchers had decided that they would do so by gathering saliva samples from Lawrence residents. Given the very grave concerns in Lawrence about high levels of pollution and their potential health impacts, the researchers simply assumed that community residents would provide any data needed. The team, when questioned, was unprepared to answer the community’s questions, such as why this research approach was selected, who would receive the results, who would own the data, and...
what would be done with the residents’ saliva samples. In the face of so many unanswered questions, many within the community became concerned and unwilling to cooperate.1

With good reason, communities that have a history of being “researched out” are shutting their doors and local government officials are taking note. In response, municipalities, like the above-mentioned community, have partnered with local academic institutions to create research guidelines for “outsiders” to follow in addition to creating a municipal internal review board (IRB).

This particular type of research is known as Community Based Participatory Research (CBPR). CBPR is the idea of performing research with the community, as opposed to just on the community. Historically, academic researchers have performed experiments on a community with the hope of yielding a finding without placing the information in context and never explaining to the community the implications of the findings. Thereby, members of the community are only involved in the data-collection aspect of the research process.

In the conventional model of research, academic researchers have a hypothesis that they wish to explore. They begin this exploration by doing an exhaustive literature review and formulating a process for control of independent variables, in the hope of confirming the research hypothesis without mitigation or question. Once the researchers have shaped the protocol, they seek a specific population from which to gather data. To do so, they may decide to reach out to different organizations to find data providers. After the researchers have gathered data, the process of analyzing ensues, perhaps followed by publication. The data providers are often left in the dark as to the rationale of the research and the potential impact of the research.

In CBPR, academic researchers partner with and involve members of the community throughout the entire research process—from the formulation of the research question being examined—to the policy outcomes. It is different from the traditional research process, in that the hypothesis is shaped with input from the community and is typically defined by a problem faced by the community. Every step, from writing a grant to conduct the research, defining the population, determining how the data will be gathered and analyzed, where the data is stored and who has access to it, and the outcomes of the findings, involve members of the targeted data providers. Though there are challenges with this research methodology (e.g., it is very time consuming, the transient nature of community members, disparate background knowledge), it is a powerful tool in lessening health disparities as well as other societal problems.

In New York City (NYC), collaboration among community members that included residents, advocacy groups, the NYC Department of Health, and the New York Academy of Medicine tackled the increasing rate of recidivism among inmates whose homes are in Central and East Harlem. Using secondary data analysis, literature reviews and focus groups, the collaborative was able to identify policies that contributed to the increase in recidivism as well as secure funding for programs that would combat the increase.2

What does this have to do with research libraries, or more importantly, the research librarian?

As research librarians, we are involved in the research process on multiple levels. We are approached with the topic and hypothesis and asked to help develop a bibliography of the relevant research, find organizations or researchers who are engaged in similar research, and are asked to provide research instruction. Moreover some of us are involved in the organization and are charged with finding potential storage areas for the research data. The conventional model of our craft was one of a gatekeeper; however, now with the democratization of the information, we are charged with ensuring that the researchers’ information is not only credible, but contains richness that enables effectiveness and efficiency.

With CBPR, the research librarian is no longer accountable to just the academic researcher but also to the larger community as this type of engagement pushes us farther away from the stacks. Because CBPR bears all the components of successful research—collaboration, reuse of data, partnerships, and transparency—more researchers are resorting to it (if the targeted populations are not demanding it). Eventually CBPR may become a mandated component of social and health sciences curriculum at universities, calling for an increase in cultural competence and multilingualism. Culturally competent, multilingual research librarians will be of utmost importance to the academy, and community members. They will be the trust-brokers between the two entities as they endeavor to create research that is honest, contextual and complete. The research on CBPR indicates that, while it is a long process, it is effective. Additionally, Clinical Trans-
ational and Science Centers are calling for more community engagement. Look back at the example of the Lawrence saliva research project presented in the introduction. Imagine for a minute that, within that academic research group, there had existed a research librarian familiar with the Lawrence community or one like it. What kind of knowledge could a culturally competent, multilingual librarian provide for the academic researchers in order to answer some of their questions about the community? What kind of instruction or knowledge could the research librarian have provided the community with in order to help them achieve an understanding of the project?

Many of the community based studies utilize secondary data analysis, such as that conducted by the government, because of its cost-effectiveness. The government has funds to collect large amounts of data with specialized tools; additionally the data used comes from the state and federal government, which makes it difficult for private industries and policymakers to question its credibility. As CBPR continues to become a norm in the research practice, accessibility (through metadata and preservation) of datasets will become an important issue.

Dr. Barboza told her audience, consisting of a large group of academic researchers, to not simply rely on others’ representations of data, but rather to collaborate with community partners to find true, if not new, meaning in the raw data than what might be presented in traditional analysis. As research librarians we have a role to play in this paradigm, but like the conversations we have when wanting to partner with scientists, we are asked to observe, learn and listen from scientists in hopes of developing a competencies that will allow for a fruitful partnership. CBPR requires that we develop relationships that are accountable to all shareholders in the research and developing that relationship requires cultural competencies, multilingualism, understanding of inequities and the historically impact of research on those who may have lacked, to some degree, agency over their bodies. Hopefully as we move forward with research initiatives we take into account that this requires more than just symbolic representation.

Endnotes


Serving the Future E-Science Researchers
Library Computing Data Services for Graduate Students on the Rutgers Newark Campus

This is an era when both the research world and library world are experiencing big changes. In the research world, networked data intensive research is fast becoming the new scientific paradigm, and librarians are trying to partner with researchers deeper into their research process to address this new paradigm. As the researchers are using and sharing more and more data, libraries are working to provide data curation, access, and computing support; thus creating a new niche of library data services.

At Rutgers University, the main campus library has taken charge of the institutional data curation responsibilities, and the John Cotton Dana Library at Rutgers’ Newark Campus, one of the Rutgers University branch libraries, has many experienced reference librarians who help users identify and access research data. To address the additional need for data computing, the Dana Library has created a new data services librarian position.

Since becoming the data services librarian, several research centers and institutional departments on the campus have approached the library to express their data computing needs. A continuous conversation is going on between the centers and

Minglu Wang is the data services librarian at the John Cotton Dana Library of Rutgers University. She earned her BA and MA in Sociology from Fudan University in China, and MLIS from the University of Iowa. She was a ARL 2009 CEP fellow and is currently providing data services at Rutgers, focusing on data analysis and presentation.
I created an additional section comparing different software in terms of their particular advantages and disadvantages in order to help users choose the appropriate tool for their own projects.

In these introductory workshops, I included best practices of doing social science data analysis, especially on issues of documenting the analysis process by using program syntax file, rather than merely using the point-and-click menu options. I also discuss the importance of doing research within the mindset of replicability of analysis, because it makes their work efficient, especially in a team based collaborative academic environment, and it makes their work honest and potentially testable in long-term scientific procedures in a broader community of researchers. The new e-scientific paradigm needs not only more open data, but also a new research culture and practice that makes deep level of research sharable and replicable.

Contextualizing data analysis within a broader process of research life cycle is another feature that I have been trying to integrate into my introductory workshops on statistical software packages. Through my experiences working with students who brought their concrete data analysis difficulties to me, I realized that many problems could have been avoided if they had a better design of their research from the very beginning, even before they began collecting data. Graduate students have difficulties and are guideless about how to organize and clean the raw collected data in a way that turns it into a ready-to-analyze standard dataset. Addressing these issues requires graduate students to have a broader view of the research process.

In order to better understand students’ specific interests and disciplines, I sent an online survey via email to the entire graduate student body. Within just a few days, more than 160 students responded. Of these responses, 142 showed interest in library workshops on using data analysis software. Interestingly, the results of the survey showed that individual students and departments/schools have unique preferences towards the kinds of software they want to learn more about. For example, Business School students are more interested in learning SAS (Statistical Analysis System); students from the School of Public Affairs and Administration are more interested in Stata, and students from departments in the Graduate School sought training in SPSS (Statistical Package for the Social Sciences).

Encouraged by the survey results, I started to offer workshops on using data analysis software SPSS, Stata, and SAS. Some students even signed up to learn all the packages. In response, the library to try to define our role and capacity to help them with their needs. Although the needs of the centers and departments are being addressed, it is more challenging to address the data analysis needs of graduate students who are not as aware of the new services provided by library and whose data analysis needs are not specifically voiced to the library. My expertise in multiple statistical software packages has helped me to get to know the graduate students’ data analysis needs and to help them through different venues. When working with these students, I always keep in mind that these are the future e-researchers, who are the main players in a new academic world where networked data intensive practices are the normal way of conducting research. With this is mind I provide more than just technical support, which is usually offered by the university’s lab or computing services departments.

The Rutgers-Newark campus computing labs are equipped with multiple statistical software packages; instructional support for using these tools is, however, not available other than during students’ research method classes, where the main focus of the class is statistics or research methodology and not how to use the software. When the library learned that other universities had started offering introductory classes on these data analysis packages; we decided to see if starting similar services at Rutgers was something students would be interested in.

The librarian’s role as educator has now extended into that of research partner, and within this new role librarians are quickly learning how to best serve the needs of users in this evolving digital and e-science academic world.

In my workshops, I recommended guidelines on preparing research data in case they need it to process their current data and plan for future projects. I stress the importance of the holistic view of doing research: having a better plan at the beginning with the future analysis and possible result in mind. Storing and organizing research data and analysis files are other problematic areas that students seldom learn from their classes. Despite this fact, I have not heard concerns about data file management and archiving their data for the long-term preservation and sharing. It is quite understandable since they are in their early stages as researchers and because even their faculty have just started to learn about these issues or may not even realized that these problems...
exist. The library world is by far the most enthusiastic group about advocating for data management best practices, and several of the leading library data services have been developing valuable guides for researchers. In response I have participated in maintaining a library guide on the topic with the data services librarian and will present it in all the workshops that I teach. Student feedback from the workshops has been positive, and they appreciate the resources in our library guides, although some don’t know yet whether these are going to be practically useful for them in the short-term.

The librarian’s role as educator has now extended into that of research partner, and within this new role librarians are quickly learning how to best serve the needs of users in this evolving digital and e-science academic world. Data services librarians are essential to providing services that address the new needs and issues related to all aspects of data based research processes. As this new function within libraries emerges we need to continually learn about data services needs, to adapt to users whose expressed needs are often vague and even unrealized. It is so exciting to be able to lead users, especially the future researchers, to think more broadly than their current project’s concerns, to develop good documentation habits, and to be prepared for a new research environment.

Endnotes


