Purdue University Purdue e-Pubs

Libraries Research Publications

2009

A Subject Librarian's Guide to Collaborating on e-Science Projects

Jeremy R. Garritano *Purdue University,* jgarrita@umd.edu

Jake R. Carlson *Purdue University,* jakecar@umich.edu

Follow this and additional works at: http://docs.lib.purdue.edu/lib_research Part of the Library and Information Science Commons

Garritano, Jeremy R. and Carlson, Jake R., "A Subject Librarian's Guide to Collaborating on e-Science Projects" (2009). *Libraries Research Publications*. Paper 140. http://docs.lib.purdue.edu/lib_research/140

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

A Subject Librarian's Guide to Collaborating on e-Science Projects

Jeremy R. Garritano* Acting Head, Chemistry Library Chemical Information Specialist Assistant Professor of Library Science M. G. Mellon Library of Chemistry Purdue University

Email: jgarrita@purdue.edu Phone: (765) 496-7279 Jake R. Carlson Data Research Scientist Purdue University Libraries Purdue University

Email: jrcarlso@purdue.edu Phone: (765) 494-6665

*Author to whom correspondence should be sent.

Abstract

For liaison or subject librarians, entering into the emerging area of providing researchers with data services or partnering with them on cyberinfrastructure projects can be a daunting task. This article will provide some advice as to what to expect and how providing data services can be folded into other liaison duties. New skills for librarians and traditional skills that can be adapted to data curation work will also be discussed. A case study on the authors' experiences collaborating with two chemistry faculty on an e-science project serves as the framework for the majority of this article.

Introduction

E-science has been defined in a variety of ways, depending on the perspective of the person defining it. Hey and Hey (2006), in an article written on the implications of e-science to the library community simply defines it as "networked, data-driven science." The implications of such a concise statement are far-reaching across many disciplines, not just science (Commission on Cyberinfrastructure for the Humanities and Social Sciences 2006). Over the last several years, documents have been published that introduce librarians to the emerging areas of data curation, cyberinfrastructure and e-science (Gold 2007a, b; Jones 2008). The e-science paradigm of research is growing as tools and resources to support its practice are becoming more widely available and funding agencies are making e-science a priority (Blue-Ribbon Advisory Panel on Cyberinfrastructure 2003; Joint Information Systems Committee 2009; National Science Board 2005). There is currently a great deal of speculation on the role of libraries and librarians in supporting e-science. It has been suggested that librarians could provide leadership in such areas as digital scholarship, open access and open data; digital stewardship and preservation; metadata standards and creation; and training future data scientists or librarians with the skills needed to successfully contribute to e-science (Case 2008; Joint Task Force on Library Support for E-Science 2007; Messerschmitt 2003; Swan & Brown 2008).

However, given only these introductory articles or overarching reports on e-science, what are librarians to do when their library administration makes a push for librarians to participate in e-science? Or what if they are simply approached by a group of researchers for assistance with

their data needs? What does a librarian bring to the table? What does one need to do to be able to contribute? What skills or techniques possessed by librarians can readily be applied to the e-science environment? What additional skills do librarians need to learn? At the very least, what can a librarian expect to encounter once they become involved with an e-science project?

With a few exceptions, the literature has been silent on these more practical aspects. Choudhury (2008) indicates libraries and librarians can become a "trusted, objective," and "verifying authority" when it comes to data and can also support scholarly communication, especially related to how it affects collection development. Steinhart (2006) discusses the role of libraries in the need for communication among interested parties related to data-sharing agreements and policies. Others touch on the need for librarians to understand data repository software and to assist in the creation of "simple, clear metadata standards" (Allard et. al. 2005). Overall, these communications are still closely tied to the concept of an Institutional Repository (IR) or digital library, and do not touch on skills needed by librarians related to other practical aspects such as relationship building and the pursuit of funding if they are to fully participate in e-science. On the other hand, some discussions do not even include the term "librarian" at all (Borgman et. al. 2007) and therefore if there is a role for librarians to play in e-science, they must continue to build their skills so they can adequately contribute and be recognized for these same skills.

The authors have ventured into supporting the data needs related to an e-science project and have identified five categories of skill sets that librarians new to this area should expect to adapt or develop when participating on such projects:

- Library and information science expertise
- Subject expertise
- Partnerships and outreach (both internal and external)
- Participating in sponsored research
- Balancing workload

Background at Purdue University Libraries

The Purdue Libraries began to take steps toward a new research-oriented conceptualization in the early 2000's, with the creation of a new series of subject librarian positions designated "information specialists." As more library resources became accessible over the Internet and interdisciplinary research became a further focus of the University, the information needs of the faculty, staff, and students at Purdue became less centered by academic departments and more focused on subject knowledge. For example, although connected to the Chemistry Library, the Chemical Information Specialist reaches out to and works with those who need help with chemical information specialists have few responsibilities in managing a physical library and are more detached from its location, it became easier for them to become connected with multiple departments and research groups. The Purdue University Libraries currently have Information Specialists for chemical information, geographic information systems (GIS), business information, agricultural information, and biomedical information. Since these positions focus on outreach and making connections across the University, they became natural positions to

leverage and further adapt as the Libraries shifted their strategic directions to become more involved in interdisciplinary research.

Librarians, even information specialists, at Purdue University did not begin supporting e-science overnight. First, with the arrival of a new dean in 2004 there began a stronger emphasis placed on research, and more specifically, interdisciplinary research. Librarians at Purdue have faculty-status and are expected to participate in research to fulfill the requirements of promotion and tenure. To support the data needs of other Purdue researchers however, the new dean believed that the Libraries would have to go one step further and become a research unit in its own right.

Transforming the Purdue Libraries into an academic research unit required the clear articulation of expectations for research from the Libraries' faculty. The Libraries' 2006 strategic plan broadly outlined these expectations. The preamble set the stage in saying that "The Libraries faculty are committed to advancing knowledge through disciplinary and interdisciplinary research." The meaning of these words was further articulated in the Libraries mission statement: "The Libraries faculty and staff are grounded in the principles and practices of library and information science... They bring their library science expertise to collaborative initiatives with colleagues in other fields to more effectively undertake interdisciplinary research..." Metrics included the number of collaborative grant proposals, the number of Libraries faculty participating in collaborative or sponsored research, the number of presentations at conferences and publications resulting from interdisciplinary research, and the amount of new funding allocated for interdisciplinary research (Purdue University Libraries 2006).

To fulfill the sentiments of these statements, the Libraries administration would need to devote resources and undergo some reorganization to assist and support librarians with their interdisciplinary efforts. Specifically, a Libraries Research Department was created in 2005, directed by the newly created position of Associate Dean for Research (ADR). Under the auspices of the Research Department, an Interdisciplinary Research Librarian and a Data Research Scientist were also hired. The authors and others from Purdue have previously published or presented in more detail about the reasoning and exact nature of these reorganizations (Brandt 2007; Carlson & Garritano In press; Mullins 2007).

Expectations for research are also being reflected in new library positions being created at Purdue. All job advertisements and job descriptions for new faculty or professional hires in the Libraries include a statement to the effect that engaging in interdisciplinary research is a requirement of the position. The exact nature of the statement varies according to the nature of the position, but can state something similar to, "willingness to participate in interdisciplinary research initiatives." In some positions interdisciplinary research is only one of several requirements, while in others it is the primary focus of the position.

Case Study: Purdue Librarians and CASPiE

The authors have worked jointly on an e-science project at Purdue and this experience has led to the realization of the five categories of skill sets listed earlier. This section will introduce the case study while further details of the case study will be included in the appropriate skill areas discussed afterward.

The Center for Authentic Science Practice in Education (CASPiE), is a multi-institutional, National Science Foundation (NSF) funded, undergraduate research center that is centrally administrated from Purdue University's Chemistry Department (CASPiE 2004). One of the primary goals of CASPiE is to provide authentic research opportunities in the first and secondyear college curriculum. These opportunities are created by a faculty member who creates a course module centering on an actual research problem the faculty member is pursuing. The multi-institutional nature of CASPiE allows the faculty member to benefit from the higher capacity of experimentation due to many undergraduates looking at the same research area and from the students using an associated online instrumentation network. This network grants CASPiE participants the use of a variety of scientific instruments that their institutions might not normally have access to due to cost, maintenance, or the impracticality of providing enough instruments for all students involved. CASPiE students are able to send samples to the instrument's location and then run the instrument remotely via the internet. Information Technology at Purdue (ITaP), the University's information technology unit, assists CASPiE administration in setting up the network and manages how the instruments are accessed and used. Students are exposed to research-quality instruments and are able to generate researchquality data as a result (Lytle et. al. 2008). Ultimately, it is hoped that the results obtained from students will enable the faculty member who wrote the module to publish the outcomes of interest while giving the appropriate students credit.

This last statement is loaded with potential data management issues: How will the data be collected, managed and organized especially across multiple institutions? How will it be described and delivered to the module author? How are proper attributions maintained as to what students generated what pieces of data?

The authors' involvement with CASPiE began in April 2007 after attending a seminar presented by CASPiE's Director of Instrumentation Networking entitled, "Making Instruments Part of the Cyberinfrastructure." While much of the seminar was technical in nature, the Director also acknowledged issues related to data management as CASPiE expanded across multiple institutions. After the seminar the authors decided to contact the Director of Instrumentation Networking to express an interest in meeting to further discuss the data needs of CASPiE. The Director responded indicating his interest in learning more about what the Libraries could do for CASPiE and a meeting was set-up with him and other CASPiE IT personnel to learn more about the instrumentation network and how it is used by students. A follow-up meeting further clarified logistical aspects related to scheduling student access and the security of the instrumentation network. However, additional meetings were needed with other CASPiE stakeholders because no one person or position could speak to all of the data management needs of the program.

From these series of meetings, it was clear that CASPiE staff had several needs that could be addressed by library expertise. A proposal for a pilot project was drafted as means to demonstrate how the Purdue Libraries might begin to help CASPiE and it was submitted to the Director of Instrumentation Networking. The Libraries pilot project proposal was accepted after some revisions. The initial proposal was to provide 200 hours of Libraries staff time to build a

prototype workflow to disseminate and archive the data for one of the CASPiE modules. Specifically, the Libraries would:

- Determine needs for access/preservation;
- Inventory data and determine appropriate manners of description (i.e. metadata);
- Outline the scientific workflow and map it to data curation functions; and
- Document the process and challenges faced.

For many researchers, the only way to conduct large scale e-science is to acquire funding in order to hire staff, procure the necessary equipment, and pay for other expenses. The hope of the Libraries was that after the 200 hours of staff time devoted to the project, the CASPiE administration would see the value and expertise the Libraries could bring to CASPiE and joint funding could be sought to pursue implementing the prototype into a more fully functional workflow. Because the Libraries demonstrated their added-value, during the late fall of 2007 and winter of 2008, the authors were involved with collaborating with CASPiE and ITaP on an NSF grant proposal.

However, to be successful, the five skills sets previously mentioned would come into play as the authors would not only have to apply their library science and subject expertise, but develop relationships internal and external to Purdue that would eventually lead to the collaboration on the grant proposal.

Library and Information Science Expertise

A librarian's training in library science enables him or her to see the bigger picture with data curation issues. Skills developed in reference work, information organization and management, and collection development are crucial to successful participation on e-science projects.

The classic reference interview, used to identify and interpret the information needs of patrons, still has an important place when dealing with researchers that produce and work with large amounts of data. To help develop a data management plan or provide support for an e-science project, it may be necessary to know, for example:

- What kinds of data are being generated (formats, amount, etc.)?
- Who should have access to the data beyond the initial researcher(s)?
- Who owns the data?
- Should any restrictions be placed on the data?
- How could the data be used, reused, and repurposed?

Many of these issues can be identified through a series of data interviews, similar to the more traditional reference interview. While no standard set of questions have been developed, some have suggested key points that could be addressed during an information gathering session such as a data interview (Carlson & Witt 2007).

Data gathered during these interviews can then better inform the librarian as to the proper organization and management of the data generated by the researcher(s). Librarians are trained

to understand systems and methods of organizing disparate types of information in ways that enable discovery and access. They are also trained to arrange and disseminate materials to specialized as well as general audiences. While call numbers won't be assigned to data so it can be shelved in neat little rows, the data can still be arranged in a logical manner whether by subject, experiment, or other hierarchy worked out with the researcher(s). Additionally, the metadata used to organize the data will also play a key role in how that data will be discovered and repurposed, especially by those unfamiliar with the initial project. A librarian can also help with data management in terms of security, working out any embargos with the researchers, or creating different levels of user access to the data.

Finally, as a data organization and management plan is developed, librarians also incorporate their knowledge of collection development. Just as librarians read book reviews, view book slips, and sift through purchase requests from faculty and students to build a collection of information resources to suit the needs of their patrons, a similar process is necessary when dealing with data. Broadly speaking, data is simply another information resource, much like a book or journal, but sometimes without the more clearly defined "packaging" one might encounter between two covers or easily held in a DVD or CD-ROM slipcase. As is the case with other types of information resources, not every single piece of data may need to be added to a collection. The important portions of the data need to be ascertained, then extracted, labeled/tagged, and stored for future use and preservation. This might mean sifting through a single text file and identifying the specific portions or lines of data that need to be kept while the rest can be discarded. It might mean that keeping only the raw data is necessary because it can be further manipulated with other discovery tools and one does not need to retain the heavily processed data because it can always be replicated. On the other hand, it might be important to keep the manipulated data in order to maintain the integrity of how the original data was changed. In the end, a librarian can help researchers build their collections of data in order to make it as useful as possible to current and future researchers who may be interested in the same or similar research.

From the initial meetings with CASPiE it was learned that in the CASPiE module selected for the pilot project each lab group generated its own data via a networked HPLC (high-performance liquid chromatography) instrument while also generating additional data through in-lab experiments. Outputs from the HPLC instrument included text in the form of instrument settings and a table of x- and y-coordinates, along with a graph of the data that was subsequently smoothed and analyzed by the students using the instrument's built-in software. The data would have to be organized in such a way as to maintain the provenance of the data in order to give credit to the proper students if any of the data were to be used in a publication by the module author. It would also need to be organized so that the module author could filter through the results quickly and efficiently, en masse, in order to find any significant results. On the other hand, in terms of collection development, there would need to be a mechanism where the module author could assist the Libraries in distinguishing data that should be kept for the long-term from data that had little value after the students had completed the requirements of their lab coursework (writing a lab report, grading by the instructor or teaching assistant, etc.). Not all paths of research pursued by the students would be fruitful, so all of the data would not need to be retained. In contrast to this, the CASPiE administration expressed interest in maintaining at

least a rough databank of hypotheses already pursued by students to keep duplication of research efforts across courses to a minimum.

Throughout these conversations and data interviews, it was clear that the librarians involved could bring a unique perspective, a "librarian perspective," to the CASPiE project by helping to bring together the wide range of views and needs encountered within the project. When dealing with CASPiE, the authors initially interacted with the technical staff, identifying instrument outputs, looking at data access, discussing size and growth of the data, etc. However, the technical staff could not speak directly to the usefulness of the data, how long it should be kept, who should have access to it, and so on. Therefore, the authors met with the more "educational" side of CASPiE. By speaking with a module author and the Director of CASPiE, these and other concerns were addressed and the librarians were able to help both sides come together regarding the joint technical and educational needs for the data.

An opportunity to showcase this librarian perspective occurred when staff from the NSF made a planned site visit to review the CASPiE program. This site visit included a poster session to highlight the progress made by CASPiE and articulate future directions for the program. The CASPiE administration invited the authors to present a poster of their efforts up to that point. This poster outlined the work undertaken by the Libraries in data curation and preservation, provided a rationale for archiving CASPiE data, and discussed the issues and challenges faced in working with CASPiE's data. It also contained an early version of a data management model based upon CASPiE's established workflows. The illustration of the model demonstrated the proposed data curation lifecycle for CASPiE courses, from capturing the data and needed metadata from students throughout the course, to providing these data sets to the researcher for his or her selection and input, to ingesting the data into the Libraries' e-data repository (Carlson & Garritano 2007). Presenting the poster gave the authors the opportunity to make connections with program officers from the NSF and to reinforce the idea of librarians as partners in the research process. It was also beneficial for the CASPiE administration to demonstrate to their funding agency that they were taking the issue of data management seriously and were looking to outside experts to assist them with their concerns.

Subject Expertise

Because working on an e-science project can be more intense, detailed, and long-term than a simple reference transaction or teaching an hour long class, librarians with at least some subject knowledge will benefit by having the capability to further embed themselves within the project.

A background in the appropriate discipline can help greatly with communication while working on an e-science project. Typically a librarian might help a researcher track down articles or references to assist in the pursuit of his or her research, but it is often not necessary that the librarian comprehend the exact experiments being conducted by that same researcher in the lab. When helping with an e-science project however, it is crucial that as much of the process and workflow is understood. This is where the knowledge and expertise of the subject librarian comes into play. Subject librarians are generally able speak the same language as researchers in their subject. Not only does this improve communication because all parties are on the same page, but it also allows the subject librarian to build the trust necessary to become an integrated part of the research team.

This deeper knowledge of the subject matter can also help the librarian bring other library colleagues up to speed. The subject librarian can explain information gathered in the data interview or elsewhere to other librarians who might be better versed in the technical, metadata or other aspects of the project, but still need to understand the research at a certain disciplinary level so they can better understand the needs of all parties involved with the project. This can be accomplished not only by straightforward conversations between the subject librarian and other librarians involved in the project, but could also be achieved by providing supplemental readings for the other library staff involved. Suggested supplemental readings might include basic reference sources, review articles, or representative articles published by the researcher in question.

One of the key steps in developing a data management plan for CASPiE was to map out the entire workflow students carried out for a particular lab module. Simply comprehending the terminologies used in relation to the analytical techniques employed was important in order to converse with the module author and instrument technicians involved. Therefore, the Chemical Information Specialist found secondary sources (encyclopedias and handbooks) to help the Data Research Scientist understand what a HPLC does, reasons for using it, limitations in the analytical method, and typical data generated. A solid understanding of how an HPLC works helped inform both librarians of the potential variables they might encounter in designing a prototype workflow to collect, disseminate and preserve the data from the experiment. Understanding the module itself helped them learn what particular variables affect the outcome of the experiments, including which variables the student would be able to control or modify by themselves in the lab. Ultimately, there were three concerns: understanding the techniques used, understanding what the module was asking the students to do, and understanding the needs of the module author. These three issues were all intertwined, yet needed to be independently understood by the librarians involved. Familiarity with these issues would assure the most essential data was collected and would create a more fully-functional model for a data repository.

Traditional librarian skills of current awareness and reference as they pertain to the ability to solve needs by tracking down information are important and are further enhanced by subject expertise. Librarians wanting to participate in e-science must stay up-to-date on various data formats, metadata standards, and other related technical standards in various disciplines. It is important to have an idea of which ones are gaining favor or becoming accepted formats, while also being aware of which ones lack support or have been superseded. A librarian can bring to the table the ability to find, review, and synthesize various data formats, metadata schemas, or data transfer and storage protocols across or within various subject areas. While the traditional subject librarian may or may not have much knowledge in this often technical area, other library staff may be able to fulfill these needs when working on a project, echoing the traditional librarian skill of the referral. If one cannot completely fulfill the information need, it is important to connect the researcher with the people or resources that can.

A subject librarian should also be aware of the scholarly communication trends within the discipline(s) they serve. This includes general acceptance of open access and the availability of

open or closed data repositories for data submission which may exist outside the institution. Because each discipline, sub-discipline, and even research group currently views data sharing and access in a different light, the subject librarian can become an intermediary between researchers as well as technologists involved in the project. Furthermore, librarians bring in a broader, more service-oriented perspective to e-science projects and can help anticipate and address the information needs of potential users of the project's outputs.

In the Libraries work with CASPiE's data the Minimum Information About a Proteomics Experiment (MIAPE) metadata schema was chosen to describe the student- and machinegenerated HPLC data from the instrumentation network (Sample Processing Working Group 2006). The iteration of MIAPE specifically designated for experiments using Column Chromatography was particularly targeted (Jones et. al. 2008). As its name implies, MIAPE is intended to define the minimum amount of information needed to understand experiments, including where the sample came from and how the analysis of the samples was performed, and is specifically designed for use in public data repositories. Further, the MIAPE standard has additional iterations for other analytical techniques such as electrophoresis and mass spectrometry, increasing its adaptability for other CASPiE modules. The authors thought the MIAPE standard would serve as a solid initial foundation for describing and documenting CASPiE data and would provide enough flexibility to shift to a different metadata standard should the need arise. After a proper metadata schema was identified, the authors next worked step-wise through the module to match up data generated by the students and the HPLC to the specific metadata fields of the MIAPE standard. During the course of working on the project it was discovered that the MIAPE standard did not incorporate all of the needed and desired information about the data being generated from the CASPiE experiments. Therefore several additional metadata fields were suggested to ensure that this information was captured.

Partnerships and Outreach: Internal and External

Outreach is a necessary skill for the successful liaison or subject librarian. These outreach skills are just as valuable, if not more, when related to e-science projects. Some projects are quite large and may include a variety of individuals, both internal and external to the home institution. While a librarian might be accustomed to working with one or two faculty for a particular class or working with an entire department for a serials or database review, working on a more complicated e-science project can lead to many more partnerships.

Internal Partnerships and Outreach

A successful librarian has a sense of current awareness, not only in information resources, subject expertise, or trends in data management and dissemination as discussed earlier, but also of the other institutional resources that researchers are using to solve their information and data needs. A librarian should be aware of these departments, research groups, services, or other organizational entities in order to successfully collaborate with them and build relationships for existing and future projects. Inserting the idea of a librarian's or library's involvement can pay back in future requests for assistance.

Working with CASPiE required the authors to collaborate with a number of departments and staff levels across campus. This involvement was primarily accomplished through a series of meetings designed to cultivate the relationship between the Libraries and CASPiE. The approach created a situation of escalating involvement that would eventually lead to the prototype proposal and ultimately result in a grant proposal submission.

As discussed previously, the initial meetings were with the technical staff employed by the Chemistry Department. The Director of Instrumentation Networking explained how the network is designed and how students accessed the network as well as how ITaP handles the scheduling and security of the networking. The Laboratory Manager who maintains the instruments onsite at Purdue provided valuable insight on the data that actually is generated by the instruments, including the method and settings files of the instruments and examples of data generated by samples submitted by the student lab groups. Both were able to articulate the need for some sort of data repository to capture, organize, manage, and archive the data and the desire for metadata to further enhance the data. Once the technical staff had brought the authors up to speed on the technical aspects of the program, the group next brought in the Directory of CASPiE and the module author who was from the Department of Foods and Nutrition. They were able to explain the workflow of the students outside the instruments and record the data in their non-digital lab notebooks. Descriptions, actual and ideal, related to how the final data and student conclusion are transmitted to the author were also discussed.

When it came to the point when the CASPiE administration wanted to submit a grant proposal, it was important that everyone communicated not only what they could contribute, but also discuss what resources would be required to accomplish their contributions. At this point, the Associate Vice President for Information Technology of ITaP became involved because of ITaP's current collaboration with CASPiE. Because ITaP would also be contributing technical expertise to the overall grant, it was important for the authors to discuss with the Associate VP to see if resources could be shared and to make sure all parties agreed on who would be responsible for what part of the grant.

To finalize the grant proposal, Purdue also assigned the group a grant writer to coordinate the submission and a business officer to help create a final budget that met both the institution's and funding agency's requirements. The grant writer assisted in maintaining a consistent flow throughout the proposal making sure it read smoothly as a single submission. The business officer knew the standard charges for various services and also had access to salary and other financial data to create a more accurate budget.

External Partnerships and Outreach

During the course of a project, the opportunity may arise to meet with outside companies or organizations that provide data management services, create and approve metadata or technical standards, supply technical equipment or instruments with particular data outputs, or provide similar types of services or products. If such interactions can be arranged, then a librarian should take full advantage of the opportunity.

While working with CASPiE, the authors were fortunate to have had several meetings with staff from Indigo Biosystems, Inc., a company that deals with data management systems using open data format standards, particularly in the pharmaceutical and biotechnology industries (Indigo Biosystems, Inc. 2009). To provide the needed functionality to the data collected from CASPiE, it would be crucial to use a proper metadata standard. Discussions with Indigo included a review of possible candidates for metadata standards that could be applied to data generated through the CASPiE program. As a result, particular schemas were weeded out that were too general or too complicated for this project, not well supported, or still in a state of development. Because of their experience and expertise, staff members at Indigo were key in identifying the MIAPE metadata scheme discussed earlier in the "Subject Expertise" section.

When working with external partners, one must realize that they might be able to provide a service or equipment more efficiently than what a library organization can provide. Depending on the library's resources and funds and the attitudes of the researchers involved, it may be acceptable to involve these external partners to provide the needed service or equipment. There could be some trepidation in allowing external parties a foothold into an area where some say libraries should be staking out a claim and it may seem competitive to involve these other parties, but it is important to realize and balance the pros and cons of such relationships. Provided that roles and expectations are negotiated properly and are clearly understood by all parties, libraries stand to gain access to needed resources and expertise that can assist their work on a project.

When dealing with some companies or organizations, be expected to sign non-disclosure agreements before seeing the full details of a particular service or software. These are similar to agreements a librarian might sign with a database vendor before beta testing a new database or user interface. Do not be put off by such requests or requirements, although be sure to read them carefully and ask for help in understanding the terms and conditions if needed. The authors have found that just as this is an emerging field for librarianship, many of the companies or organizations one might deal with are also new. These start-ups need to protect their intellectual property if they are to survive.

Maintaining relationships with these external partners could also lead to a more appealing grant application. Not only do some granting agencies favor or require interdisciplinary cooperation within an academic institution or between academic institutions, but a stronger case can sometimes be made for including an industrial partner as well. These external partners might provide some sort of consulting service on the data management workflow, provide technical support for the initial set-up of a data management service, or customize equipment or software to fit the particular needs of the project, institution, and/or funding agency. It is also okay if no partnerships form after one, or even several, conversations. This is simply one aspect of risk taking that must be accepted when beginning to pursue projects related to e-science.

Participating in Sponsored Research

The frequent necessity of sponsored funding in order to conduct e-science can present an area of unfamiliarity for subject librarians. However, librarians should not be afraid to participate in grantsmanship. When the ability to acquire funds means the difference between initiating (or

continuing) a large scale e-science project or not, it is crucial to first understand what the funding agency is looking for and what is specifically required in the grant proposal. Communicating with program officers and others involved in grant reviewing can help clarify the goals of the agency or program and can help focus any grant proposal that is eventually submitted.

In the latter half of 2007, the CASPiE administration was monitoring for potential sources of sponsored funding to expand the CASPiE program and to improve upon its instrumentation network. A suitable grant solicitation was identified and all parties began to plan how they might contribute to an overall proposal. After some internal discussions, a NSF program officer for the grant was contacted in order to arrange a telephone conversation with him to discuss the ideas for a proposal. The program officer was able to ask questions that enabled the group to identify gaps or potential ideas for expansion that they had not yet considered. Also, by communicating with the program officer, we were able to make sure that our ideas for a proposal were in alignment with the particular solicitation.

When it comes to actually writing the grant proposal, librarians have the opportunity to contribute throughout the entire process. High-profile funding agencies such as the National Institutes of Health (NIH) and the NSF are beginning to require statements and plans for data management from the research being funded (Office of Budget, Finance & Award Management 2005; Office of Extramural Research 2003). Therefore, one area that librarians can contribute their skills and knowledge would be in helping to craft a data management plan for the grant proposal. The authors have found that researchers would much rather concentrate on what they do best (the actual science), than worry about data management and curation issues like where data is going to be stored, how it is going to be backed-up, or what metadata should be associated with the data. The National Science Foundation has specifically identified librarians as being well positioned to address these types of issues (Cyberinfrastructure Council 2007). If librarians can show they can address these tasks effectively, they may be accepted as either co-Principal Investigators or at least named personnel on grant proposals.

As librarians become more involved in supporting e-science, they will also need to be able to actively participate in all aspects of seeking out, applying for and carrying out grants. This means that librarians must become knowledgeable about the lifecycle of grants and the requirements that come with applying for and accepting funding. Applying for a grant is much more than simply writing about what one plans to do if awarded funding. First, realize that writing a grant takes a significant amount of time and often includes a flurry of activity as the deadline approaches. Also consider that the larger the proposal, the more likely that various pieces of the proposal will be coming from different parties and even different institutions as well. These pieces then have to be stitched together to form a cohesive narrative in order to strengthen the purpose and goals of the proposal. Be sure that each iteration of the proposal is carefully reviewed to be sure that all parties can stand behind it.

Second, librarians need to learn to speak the language of grantsmanship, which often revolves around staff, equipment, and other resources. It is one thing to be able to write about what the library or librarian can do for a particular project, it is another to estimate the resources needed to complete this and to develop a realistic budget to obtain these resources. Grants are usually

limited by some finite funding amount and each party contributing to the proposal must be flexible enough to fit all of their needs into the total amount being requested. That means not only do librarians need to articulate what they will do during the funding period, but they must also enumerate how many staff they will need to accomplish this as well as if any additional resources, such are hardware or software, will need to be covered by the grant funding. Very frank discussions will need to be had before a grant proposal is submitted. A university will often claim a certain percentage of the grant to cover "facilities and administration" or "F&A" costs. This needs to be taken into consideration when calculating the project's overall budget. Most universities have a research office or other unit on campus that oversees and assists faculty in submitting grant proposals. Even if the libraries are not leading the proposal it is a very good idea to make contact with the office that oversees the process of submitting a grant as they can help throughout the process and inform as to what will be expected. The library business office should be notified early on in the process as well so they can help generate a realistic budget and will likely need to work with other units in the university to pull together the information needed for one to participate in the grant. Every institution does things differently so be sure that local practices are understood.

Finally, librarians need to realize the time scale on which many grant proposals operate. Many librarians are accustomed to a semester or quarterly cycle. Classes can be framed in a semester, with a lifespan of months. A semester itself may be very busy in the beginning, slow down until mid-terms, and then become quiet again until right before finals. A collection development cycle may run from fiscal year to fiscal year. However, many grants run on a multi-year basis. That means a librarian must realize that the timeline for a grant is much further out than might usually be anticipated. Some grants can run from 3-5 years at a time, with the librarian having various levels of responsibility in each of those years. For example, this may mean the librarian is heavily involved with the initial set-up in years 1 and 2, and then comes in again during year 5 to assist with an outreach plan. Or a librarian might sign on to be on a grant but not even be involved much at all until years 4 and 5. Librarians must be prepared to make longer commitments when becoming involved with such projects and allot their time accordingly.

Because of their previous work with CASPiE, when the CASPiE administration identified a grant solicitation that interested them, the authors were asked to contribute to the proposal. The first step for the authors was to read through the solicitation to identify areas where the Libraries could contribute to and support the overall proposal. Because ITaP was also invited to partner on the grant proposal, emails were exchanged in an effort to sketch out who could contribute what and what additional resources would be needed in order to start framing a budget. A principal investigator (PI) was identified within the group to lead the project being proposed and to coordinate the grant proposal. A couple of face-to-face brainstorming meetings followed to connect each contributor's areas of expertise in order to create a cohesive, single proposal. Next the authors had to negotiate with the Libraries' administration and their supervisors as to the amount of their time they could contribute to the proposal. Most grants require a percentage of effort be placed on specific positions or named personnel. In this case, because of previous commitments and the five year time frame of the proposal, the Data Research Scientist negotiated 20% and the Chemical Information Specialist negotiated 10%. Additionally, the Libraries requested the funds to support hiring a graduate student at half time to assist the Libraries' efforts on the project.

In addition to describing the librarian's proposed efforts and developing a budget for the project, there are several possible additional tasks associated with grant proposals. First, the authors had to recommend specific people to serve on an advisory board for the resulting grant who were external to Purdue and who could speak to the areas to which the Libraries were contributing. The authors also had to obtain a letter of support from the Libraries administration indicating the importance of the role of the librarians in the overall project and more generally why the Libraries were needed as partners. Finally, the authors had to include a brief biographical statement and a short write up of each individual's research activities. These last two documents were meant to highlight the librarians' qualifications to justify to the proposal reviewers their ability to accomplish what they actually stated in the grant proposal.

Balancing Workload

A librarian might ask, what is the typical time commitment for an e-science project and how can I incorporate this level of participation in my current workload? It is difficult to answer this question because each e-science project is different from another. These projects can live or die based on availability of resources or funding, personnel who stay at or leave the academic institution, or the research is no longer an area of interest for those originally involved. When sponsored funding is involved, things can take even longer. Developing a project to the point where a grant proposal might even be considered could take months, crafting the proposal takes time, and then there is a period of months while all proposals are being reviewed before a final decision is made. After all of this, the grant might not even be funded and thus it may be perceived that all of this effort has lead to a failure.

Therefore, although there needs to be some recognition of these perceived risks when becoming involved in e-science projects, there also needs to be an acknowledgement that especially in the initial stages, any involvement by librarians is a learning experience and will build capacity. Researchers experiment and pursue funding for areas that look promising, yet some experiments fail or need to be readjusted based on initial results. The same can be said for librarian involvement in working with e-science projects. Mentally, it may be more difficult to put effort into crafting a multi-year project when there is the possibility it will fade away without any solid results. However, partial success or a perceived impact can still occur if one has accomplished such things as:

- Building relationships with researchers within their institution
- Raising the level of involvement of the library organization within interdisciplinary research
- Developing general standards, policies or protocols for data management that can be used with other projects
- Providing partial solutions to the data and information needs of local researchers

If there is still some hesitation regarding the amount of time required, a tact the authors initially took with the CASPiE administration was to define the amount of time that would be spent on the project. In this case we limited ourselves to a total of 200 staff hours to the initial investigation. Librarians would document their work and provide a report that would hopefully

demonstrate to the CASPiE administration that pursuing additional resources or funding for the Libraries' work with CASPiE would be justified. This concept was understood and accepted by CASPiE. Knowing the amount of data involved and the ongoing nature of the project, CASPiE administrators realized that the implementation of the suggested solutions could not be done for "free." Many researchers realize the only way to keep their research going is to obtain sponsored funding or grants. Though many researchers have perceived the traditional services of the Libraries as free-of-charge in the past, the authors have found that researchers understand the need for additional funding and resources for the libraries to assist them with particular aspects that are beyond their own areas of expertise.

The authors acknowledge that committing time and effort to e-science endeavors can be risky. Relationships can take longer to evolve and significant payoffs may be years off instead of months. Becoming involved in grant proposals and sponsored funding requires additional considerations in determining one's workload. Acknowledgement and support should come from one's library administration acknowledging that release time will be required if a grant is awarded and should plan accordingly. Risk-taking should be encouraged through providing recognition and rewards to even partial successes of librarian-researcher collaborations such as submitting grant proposals, even if they are not awarded. If nothing else, the production of conference presentations, articles, and other communications can be conveyed as a benefit of participating, whether a grant or project is successful or not. Despite the risks, there is a demonstrated need for information support for e-science and the kinds of organizational and information management skills that librarians possess. Through partnering with researchers, technologists and others and participating in e-science projects, librarians can help contribute to the development of a robust information ecology for this emerging area of research.

Conclusion

There are a variety of ways to initiate e-science partnerships. The authors' experiences just described related one possible entry point by becoming involved with undergraduate research initiatives. Interest in undergraduate research has expanded into many academic institutions and may be a fruitful first attempt for unsure librarians. Undergraduate research can also have well-defined technical and educational components, helping the argument for librarian involvement, as well as being less territorial for the researchers involved.

In many ways, librarians are already prepared to take the chance and become involved with escience projects. Creatively adapting library science skills and subject expertise will provide a foundation for librarians to communicate effectively with researchers and fellow information professionals to address data management needs. A librarian can also foster additional relationships and build an awareness of data management issues and problems as well as the tools and resources that exist to address these issues. Finally, pairing knowledge of grantsmanship with organizational support for e-science projects, will allow librarians to be more directly collaborative instead of simply supportive to research groups.

Through their work librarians naturally transcend disciplinary and other boundaries (technical, educational, etc.) as well as being able to identify and address the needs of both information producers and information consumers. Librarians can play a central role by using this

"transcendence" to bring people and resources together to address issues in interdisciplinary research.

References

Allard, S., Mack, T.R., & Feltner-Reichert, M. 2005. The librarian's role in institutional repositories: A content analysis of the literature. *Reference Services Review* 33(3): 325-336.

Blue-Ribbon Advisory Panel on Cyberinfrastructure. 2003. *Revolutionizing Science and Engineering through Cyberinfrastructure*. National Science Foundation. [Online]. Available: http://www.nsf.gov/od/oci/reports/atkins.pdf [Accessed February 22, 2009].

Borgman, C.L., Wallis, J.C., & Enyedy, N. 2007. Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries* 7(1/2): 17-30.

Brandt, D.S. 2007. Librarians as partners in e-research: Purdue University Libraries promote collaboration. *C&RL News* 68(6): 365-7, 96.

Carlson, J.R. & Garritano, J.R. 2007. *Preserving Undergraduate Research Data*. Purdue University. [Online]. Available: <u>http://docs.lib.purdue.edu/lib_research/82/</u> [Accessed February 22, 2009].

Carlson, J.R. & Garritano, J.R. In press. E-science, cyberinfrastructure and the changing face of scholarship: organizing for new models of research support at the Purdue University Libraries. In: Walter, S., Coleman, V., & Williams, K., editors. *The Expert Library: Staffing, Sustaining, and Advancing the Academic Library in the 21st Century*; Chicago: Association of College and Research Libraries.

Carlson, J.R. & Witt, M. 2007. *Conducting a Data Interview*. Purdue University. [Online.] Available: <u>http://docs.lib.purdue.edu/lib_research/81/</u> [Accessed February 22, 2009].

Case, M.M. 2008. Partners in knowledge creation: an expanded role for research libraries in the digital future. *Journal of Library Administration* 48(2): 141-56.

CASPiE. 2004. *The Center for Authentic Science Practice in Education*. CASPiE. [Online]. Available: <u>http://www.purdue.edu/dp/caspie/</u> [Accessed February 22, 2009].

Choudhury, G.S. 2008. The virtual observatory meets the library. *Journal of Electronic Publishing*. [Online] Available: <u>http://hdl.handle.net/2027/spo.3336451.0011.111</u> [Accessed February 22, 2009].

Commission on Cyberinfrastructure for the Humanities and Social Sciences. 2006. *Our Cultural Commonwealth*. American Council of Learned Societies. [Online]. Available: http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf [Accessed February 22, 2009].

Cyberinfrastructure Council. 2007. *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation. [Online]. Available: http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf [Accessed February 22, 2009].

Gold, A. 2007a. Cyberinfrastructure, data and libraries, part 1: a cyberinfrastructure primer for librarians. *D-Lib Magazine*. [Online]. Available: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html [Accessed February 22, 2009].

Gold, A. 2007b. Cyberinfrastructure, data and libraries, part 2: libraries and the data challenge: roles and actions for libraries." *D-Lib Magazine*. [Online]. Available: <u>http://www.dlib.org/dlib/september07/gold/09gold-pt2.html</u> [Accessed February 22, 2009].

Hey, T. & Hey J. 2006. E-science and its implications for the library community. *Library Hi Tech* 24(4): 515-28.

Indigo Biosystems, Inc. 2009. *Indigo Biosystems*. [Online]. Available: http://www.indigobio.com [Accessed February 22, 2009].

Joint Information Systems Committee. 2009. *E-Research*. Higher Education Funding Council for England. [Online]. Available: <u>http://www.jisc.ac.uk/home/whatwedo/themes/eresearch.aspx</u> [Accessed February 22, 2009].

Joint Task Force on Library Support for E-Science. 2007. Agenda for Developing E-Science in Research Libraries. Association of Research Libraries. [Online]. Available: http://www.arl.org/bm~doc/ARL_EScience_final.pdf [Accessed February 22, 2009].

Jones, A., Carroll, K., Knight, D., MacLellan, K., & Paton, N.W. 2008. *MIAPE Column Chromatography*. HUPO Proteomics Standards Initiative. [Online]. Available: <u>http://www.psidev.info/files/MIAPE_CC_1.0.pdf</u> [Accessed February 22, 2009].

Jones, E. 2008. *E-Science Talking Points for Deans and Directors*. Association of Research Libraries. [Online.] Available: <u>http://www.arl.org/bm~doc/e-science-talking-points.pdf</u> [Accessed February 22, 2009].

Lytle, F.E., Weaver, G.C., Wyss, P., Steffen, D., & Campbell, J. 2008. Making instrumentation a secure part of the cyberinfrastructure. *Unpublished Manuscript*.

Messerschmitt, D.G. 2003. *Opportunities for Research Libraries in the NSF Cyberinfrastructure Program*. Association of Research Libraries. [Online]. Available: http://www.arl.org/resources/pubs/br/br229/br229cyber.shtml [Accessed February 22, 2009].

Mullins, J.L. 2007. Enabling International Access to Scientific Data Sets: Creation of the Distributed Data Curation Center (D2C2). Purdue University. [Online]. Available: http://docs.lib.purdue.edu/lib_research/85/ [Accessed February 22, 2009]. **National Science Board**. 2005. *Long-Lived Data Collections: Enabling Research and Education in the 21st Century*. National Science Foundation. [Online]. Available: http://www.nsf.gov/pubs/2005/nsb0540/start.htm [Accessed February 22, 2009].

Office of Budget, Finance and Award Management. 2005. *Grant Policy Manual: Chapter VII - Other Grant Requirements - Section 734*. National Science Foundation. [Online]. Available: <u>http://www.nsf.gov/pubs/manuals/gpm05_131/gpm7.jsp</u> [Accessed February 22, 2009].

Office of Extramural Research. 2003. *Final NIH Statement on Sharing Research Data*. National Institutes of Health. [Online]. Available: <u>http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html</u> [Accessed February 22, 2009].

Purdue University Libraries. 2006. Strategic Plan 2006-2011. Purdue University. [Online]. Available: <u>http://www.lib.purdue.edu/admin/stratplans/plan2011.html</u> [Accessed February 22, 2009].

Sample Processing Working Group. 2006. *MIAPE: The Minimum Information About a Proteomics Experiment*. HUPO Proteomics Standards Initiative. [Online]. Available: <u>http://www.psidev.info/index.php?q=node/91</u> [Accessed February 22, 2009].

Steinhart, G. 2006. Libraries as distributors of geospatial data: data management policies as tools for managing partnerships. *Library Trends* 55(2): 264-84.

Swan, A. & Brown, S. 2008. The Skills, Role and Career Structures of Data Scientists and Curators: An Assessment of Current Practice and Future Needs. Key Perspectives Ltd. [Online]. Available:

http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinal report.pdf [Accessed February 22, 2009].