Research Paper ■

Issues in Biomedical Research Data Management and Analysis: Needs and Barriers

Nicholas R. Anderson, MS, E. Sally Lee, MS, J. Scott Brockenbrough, PhD, Mark E. Minie, PhD, Sherrilynne Fuller, PhD, James Brinkley, MD, PhD, Peter Tarczy-Hornoch, MD

Abstract **Objectives:** A. Identify the current state of data management needs of academic biomedical researchers. B. Explore their anticipated data management and analysis needs. C. Identify barriers to addressing those needs.

Design: A multimodal needs analysis was conducted using a combination of an online survey and in-depth one-on-one semi-structured interviews. Subjects were recruited via an e-mail list representing a wide range of academic biomedical researchers in the Pacific Northwest.

Measurements: The results from 286 survey respondents were used to provide triangulation of the qualitative analysis of data gathered from 15 semi-structured in-depth interviews.

Results: Three major themes were identified: 1) there continues to be widespread use of basic general-purpose applications for core data management; 2) there is broad perceived need for additional support in managing and analyzing large datasets; and 3) the barriers to acquiring currently available tools are most commonly related to financial burdens on small labs and unmet expectations of institutional support.

Conclusion: Themes identified in this study suggest that at least some common data management needs will best be served by improving access to basic level tools such that researchers can solve their own problems. Additionally, institutions and informaticians should focus on three components: 1) facilitate and encourage the use of modern data exchange models and standards, enabling researchers to leverage a common layer of interoperability and analysis; 2) improve the ability of researchers to maintain provenance of data and models as they evolve over time though tools and the leveraging of standards; and 3) develop and support information management service cores that could assist in these previous components while providing researchers with unique data analysis and information design support within a spectrum of informatics capabilities.

■ **J Am Med Inform Assoc.** 2007;14:478–488. DOI 10.1197/jamia.M2114.

Introduction

Rapid advances in analytical technology coupled with widespread access to large amounts of highly detailed, heterogeneous and often public biomedical research data have dramatically increased the difficulties faced by biomedical investigators in acquiring, archiving, annotating, and analyzing data.[1] Recognition of this fact is reflected in a number of large scale initiatives by the major U.S. funding institutions as well as a profusion of software tools designed for biomedical research data management and analysis.[1–5] Over the past several years we have met with many academic biomedical researchers to discuss solutions to their data handling problems as part of our own data integration efforts.[6–12] Through informal discussions, we have been struck by the frequency with which they stated that: a) a data handling problem had become a major barrier to the progress of their research, b) available computational solutions were prohibitively expensive, c) available solutions were too complex for their needs, and/or d) computational solutions to their problem did not exist at all. In addition, we have noticed that the needs of investigators can be extremely dynamic, often changing on a weekly basis. From a biomedical informatics standpoint, these issues raise several fundamental questions:

- How are researchers coping with managing these quickly evolving information management problems in practice?
- What obstacles are faced by researchers seeking individual solutions to data management and analysis needs?

Affiliations of the authors: Division of Biomedical and Health Informatics, Department of Medical Education and Biomedical Informatics (NRA, ESL, MEM, SF, JB, PT-H); Department of Biological Structure (JSB, JB); Health Sciences Libraries and Information Center (MEM, SF); Department of Health Services, School of Public Health and Community Medicine (SF); Department of Pediatrics (PT-H); Department of Computer Science and Engineering (JB, PT-H), University of Washington, Seattle, WA.

The authors would like to thank and acknowledge National Library of Medicine Training grant (Biomedical Health Informatics training program) T15LM07442, the BioMediator grant R01-HG02288, BISTI planning grant P20-LM007714, and the Human Brain Project grant DC02310 for providing the funding to support parts of this work.

Correspondences and reprints: Nicholas Anderson, University of Washington, Department of Medical Education and Biomedical Informatics, Box 357240, Seattle, WA 98195-7420; e-mail: <nicka@u.washington.edu>.

Received for review: 3/29/2006; accepted for publication: 3/27/2007.

- To what extent can biomedical research data handling needs be generalized across more than one lab (or even more than one project within a lab)?
- What core design issues must be addressed in designing and implementing informatics solutions to aid biomedical researchers in their data management and analysis?

To address these questions, we have embarked on a project to identify the data management and analysis needs of academic biomedical researchers at the University of Washington.

Background

Informatics journals report a steady stream of freely available analytic and archiving tools with the potential to streamline data analysis and integration tasks.[6,8,10,13,14] Yet biomedical researchers continue to struggle with increasing volume and complexity of their own datasets. Accurate and thorough needs analysis is widely recognized as one of the earliest and most crucial events in virtually all software development cycles. For example, needs analyses for a variety of applications are frequently reported in software engineering[15–20] as well in the medical fields.[21–26] However, few attempts to assess the needs of biomedical research exist in print[27] despite recent calls for increased evaluation-based science to support informatics research.[28–30] We feel that evaluation-based assessment of data management and analysis needs of biomedical research is a crucial informatics research area.

Through our examination of existing methodologies described in the literature,[21,31–37] we have concluded that mutually supportive data resulting from a mixed methods approach has the greatest potential to support a comprehensive biomedical research needs assessment. Our approach is to use broad web-based surveys followed by personalized in-depth interviews. The surveys were targeted toward a large population of biomedical researchers to provide broad overviews of generalized needs; however, surveys are limited in that they don't allow the elicitation of information that was not understood or imagined by the authors of the survey but is important to the respondents. Therefore, in addition to the quantitative survey data, we gathered highly detailed and context-specific qualitative data from individual interviews. The semi-structured interview data not only provided detailed contextual information, but helped reveal ideas that can be transferred to other domains.[17,18,21,38–40] Combining qualitative and quantitative methods has already been successfully used in the discovery of user issues associated with the implementation of clinical Electronic Medical Records (EMR) systems.[36,37,41]

In this paper, we present the results of the survey and the interviews in a combined analysis framework that we hope to use in future biomedical research needs assessments. Using this multi-modal method, we describe the needs of biomedical investigators affiliated with the University of Washington. The UW is an internationally recognized research university that was recently ranked #17 in the world by the Economist newspaper.[42] UW research supported over 7,400 full-time equivalent positions in fiscal year 2005.[43] As a result, we feel that this study, though limited to one university and its local collaborator research institutes, can be applicable to other academic biomedical research settings.

Methods

We focused our assessment on data management and analysis needs, including: a) current strategies for management and analysis of experimental data, and b) obstacles to data management and data sharing. A survey of 286 faculty, research staff, and students yielded quantifiable and moderately detailed data about informatics software needs. Fifteen volunteers from this population were the subjects of semi-structured interviews. We conducted qualitative analysis on the interview data that represented in-depth views of individual needs.

Human Subjects

To ensure the safety and anonymity of the participants, all aspects of this research including participants in both the survey and the subsequent interviews were approved by the Human Subjects Committee of the University of Washington Internal Review Board (IRB).

Survey

The survey consisted of two separate sections that together totaled 36 questions (See Appendix A, available as an on-line data supplement at www.jamia.org). Twenty-three of these questions addressed a variety of library and information science issues and built on previous UW work from Yarfitz et al. involving library-based bioinformatics services.[31] This survey is part of a process of continuous evaluation of bioresearch needs from both the academic research and institutional support perspectives. Of the 13 questions focusing on biomedical research information management needs, four questions related to subject demographics, with the remainder focusing on high-level overviews of generalized needs across biomedical research disciplines. Here we report primarily on data from the needs-related questions as well as a limited set of data from the library and information science questions that have overlap with biomedical needs. More in-depth discussion of the library-service aspects of the survey will be published elsewhere.

We pre-tested the initial survey with six volunteers actively engaged in biomedical research who were also asked to give their opinions regarding survey length and question clarity. The survey was then deployed online via WebQ, an automated survey and response analysis tool within the University of Washington Catalyst system.[44,45] Invitations to participate in the survey were sent out to 1,754 addresses in the spring of 2005. The addresses were obtained directly from an "opt-in" UW Health Sciences Center Library list[46] and were of researchers at the UW and collaborating research institutions in the Seattle area who are interested in bioinformatics resources and are actively involved in biomedical research. We estimate that the e-mail list represents approximately 30% of the active biomedical researchers in the Seattle area. The survey was left online for 7 weeks with "reminder" messages sent via e-mail every two weeks to any addresses that had not yet responded.

Interviews

Upon completion of the survey, respondents were solicited to volunteer for one-on-one interviews. A total of 15 researchers volunteered for semi-structured interviews, which were conducted in the summer of 2005. Each interview began with a "critical incident" question[38,47] followed by other relevant questions about data handling and analysis needs specific to that individual. All interviews lasted between 45 and 90

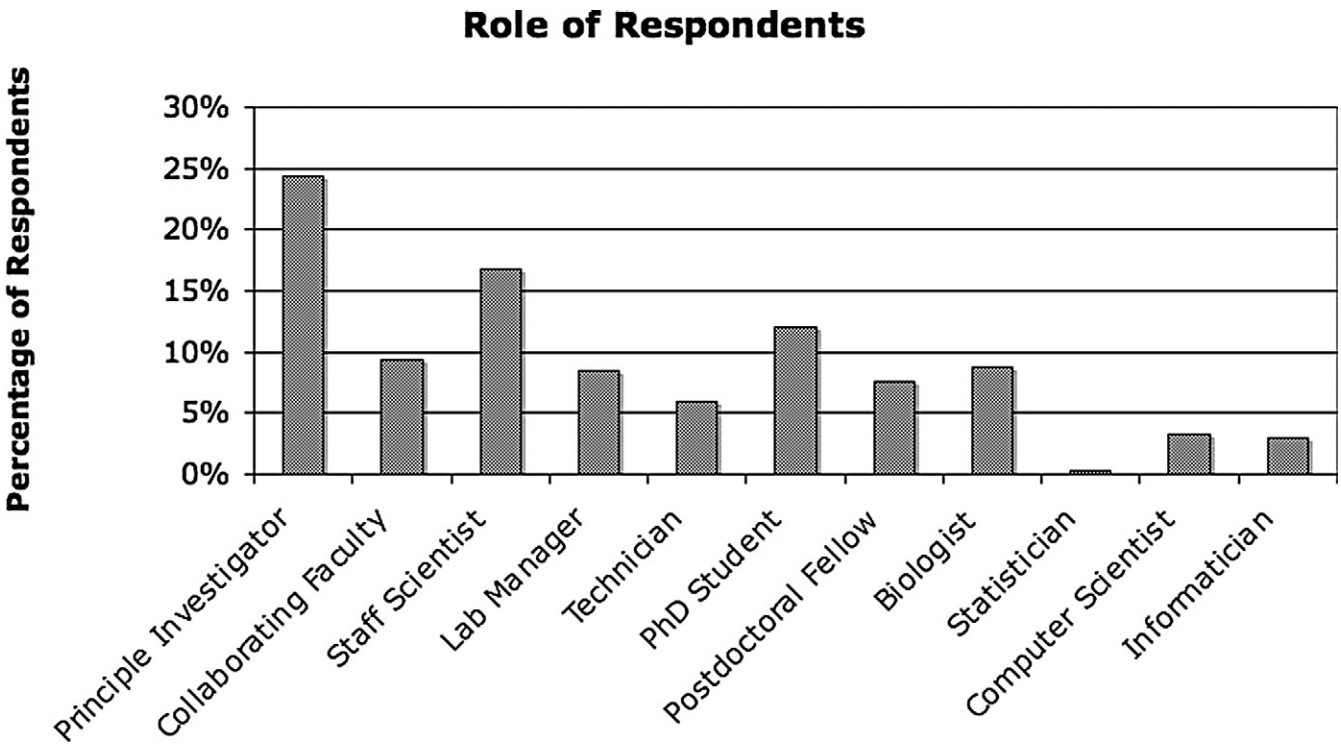

Figure 1. Primary roles by percentage of respondents.

minutes and were audio taped and transcribed with the volunteer's consent. The content of the transcripts (190 pages) were analyzed using ATLAS.Ti 5.0 software. Emerging patterns and themes were individually identified by two of the authors experienced in qualitative methods (NRA, ESL), then triangulated with input from the third author (JSB). This approach to triangulation and inter-coder reliability is consistent with other published qualitative analysis methodologies.[21] All individual themes identified were maintained separately, establishing a clear audit trail by which to resolve any potential triangulation issues.

Results

Survey Response Rate and Respondent Demographics

The survey received a total of 286 respondents, or 16% of the original list, with a margin of error of 5.3%. This is a reasonable response for a web-based survey, but lacks sufficient power for extensive statistical analysis.

Forty-six percent of respondents identified themselves as principal investigators or collaborating faculty, followed by 36% as staff scientists and lab managers, with the remainder postdoctoral fellows, students, and technicians (Fig. 1). Respondents were encouraged to select all choices that corresponded to their research roles, so the total percent figures are greater than 100%. "Small" labs (six people or fewer) and "medium" labs (7 to 15 employees) comprised 87% of respondents (44% and 43%, respectively), with "large" size labs accounting for 13% of respondents. Neurosciences, Genomics, and Cell Biology were the most frequently selected primary research sub-discipline (each specialty indicated by over 20% of respondents). Other commonly chosen areas were Biochemistry (15%), Immunology (11%), Pathology (10%), and Microbiology (10%) (Fig. 2). We examined lab size as a function of stated research specialty and found no clear association between these variables.

Needs Analysis

Three broad data management and analysis themes emerged from the analysis of the interview data within the context of the survey responses: 1) current state of data management and analysis at the laboratory level; 2) anticipated data management and analysis needs; 3) barriers to addressing those needs.

Current State of Data Management and Analysis

Eighty-four percent of survey participants indicated that they currently have or in the past had experienced data handling problems although only 52% of them sought to solve their data handling problems (Question 33) (Fig. 3). There was a clear correlation between the size of a lab and the likelihood it had experienced problems. Only 14% of survey respondents reported currently having a Laboratory Information Management System (LIMS) (Question 30).[48] When broken down by sub-discipline, developmental biologists were least likely to have a LIMS (4%) with proteomicists (18%), pathologists (17%), and cell biologists (16%) most likely to use a LIMS in their work (Table 1). Large labs were most likely to be using a LIMS (22%) but interestingly were closely followed by the smallest labs (18%) (data not shown). Most researchers (59%) were already storing at least some of their images digitally while roughly a third (34%) partly relied on hard copy archiving (Fig. 4).

Fifty percent of structural biologists and proteomicists and 48% of genomicists reported that at least 10 employee-hours per week are spent in data handling tasks (Question 29). The breakdown of weekly workload devoted to data handling as

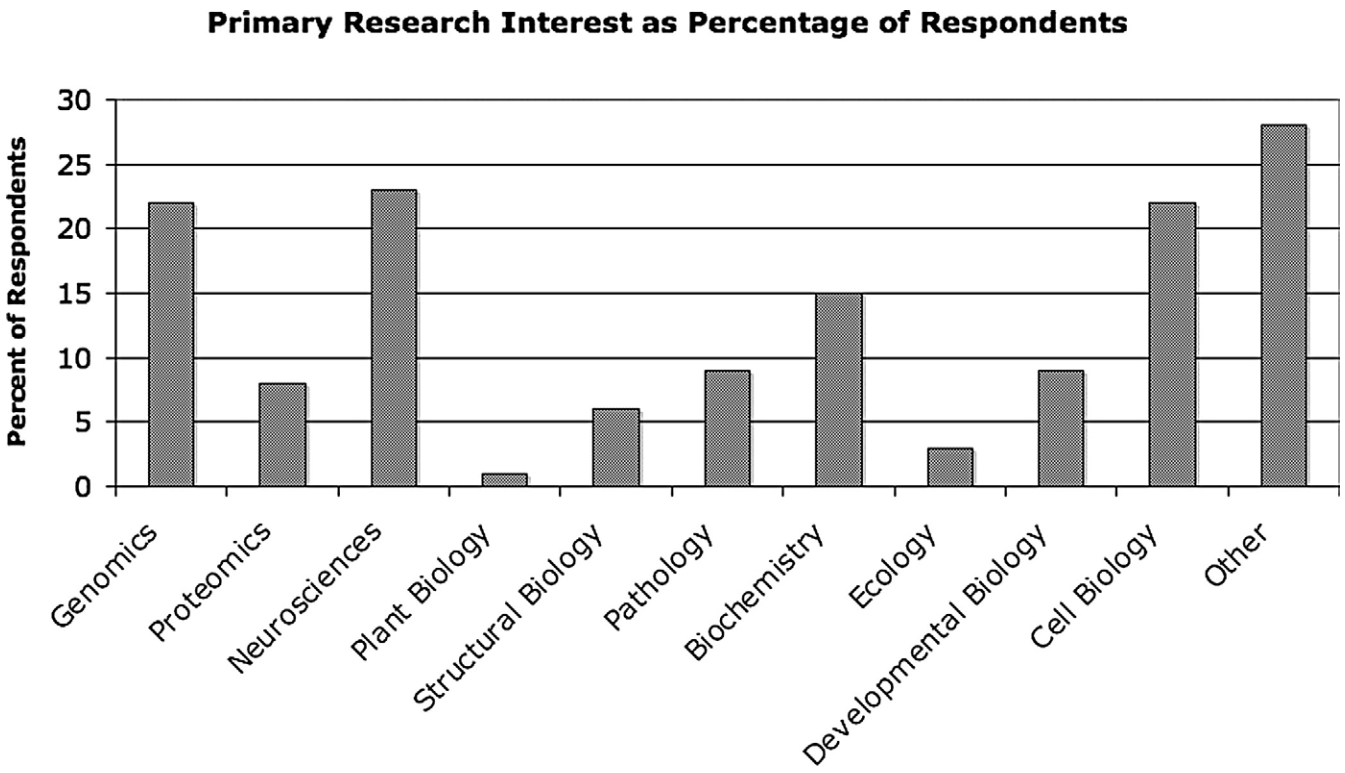

Figure 2. Primary research interest by percentage of respondents.

a function of research sub-discipline is given in Figure 5. Over 50% of survey respondents reported spending more than 5 person-hours per week in data handling tasks. Larger size labs were shown to spend more employee time each week at data handling.

During the interviews, two core themes surrounding the current state of biomedical data management and analysis emerged: the widespread use of non-specialized applications, and the difficulty of organization, storage, and retrieval of data.

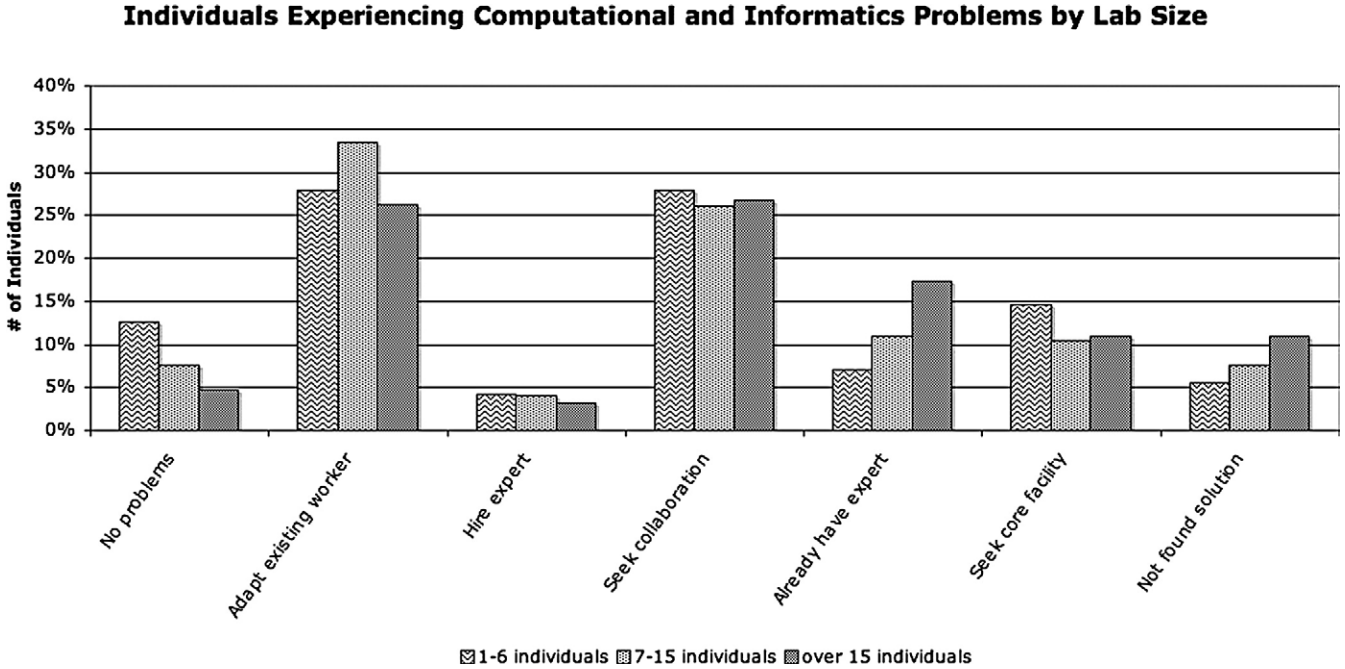

Figure 3. Individuals reporting experiencing computational and informatics problem by lab size and percentage of respondents.

Table 1 ■ Percentage of Labs Using Laboratory Information Management System (LIMS) By Sub-Discipline

Biochemistry	5.3%
Cell Biology	16.4%
Developmental Biology	4.5%
Genomics	10.7%
Neurosciences	8.5%
Pathology	16.7%
Proteomics	18.2%
Structural Biology	14.3%
All	13.8%

Common Use of Non-Specialized Applications

Most researchers already used some form of electronic organization; however, instead of using applications addressing specific needs of biomedical research, many depended on general-purpose applications such as spreadsheets, text files, and simple file sharing programs. The reasons why these tools were commonly used included simplicity of data layout, widespread availability, and short learning curve.

> "Yeah, the spreadsheet has been our main workhorse, unfortunately"
>
> "Well, that stuff I currently have just in a Word document. So, I just have it all right here (shows document on computer)"
>
> "I have one spreadsheet that has all of my chromosomes—it has a different tab page for each of 23 human chromosomes all in one Excel document - and it has all of my data it has my whole experiment and then I've gone through and color coded it for homozygosity and linkage. So this has also taken a considerable amount of time to set up, but I have this for each one of my chromosomes."
>
> "Well I'm a spreadsheet queen. So I've got everything in spreadsheets. This is just my data, I've got a spreadsheet on family information, on cell information . . ."

Nine out of fifteen interviewed researchers recognized that spreadsheet applications had disadvantages of size constraint and limited processing power. Often these researchers anticipated that they would soon reach the limits of existing general-purpose spreadsheet applications in terms of both storage and organizational capacities.

> "Well, we have multiple spreadsheets - that's one of the problems. We sort of have a master spreadsheet . . . We try to minimize it as much as we can, but I think that's a major problem."
>
> "However, that exceeds the capabilities of the spreadsheet. Spreadsheet really bogs down any time you get past say 20,000 individual cells with columns."

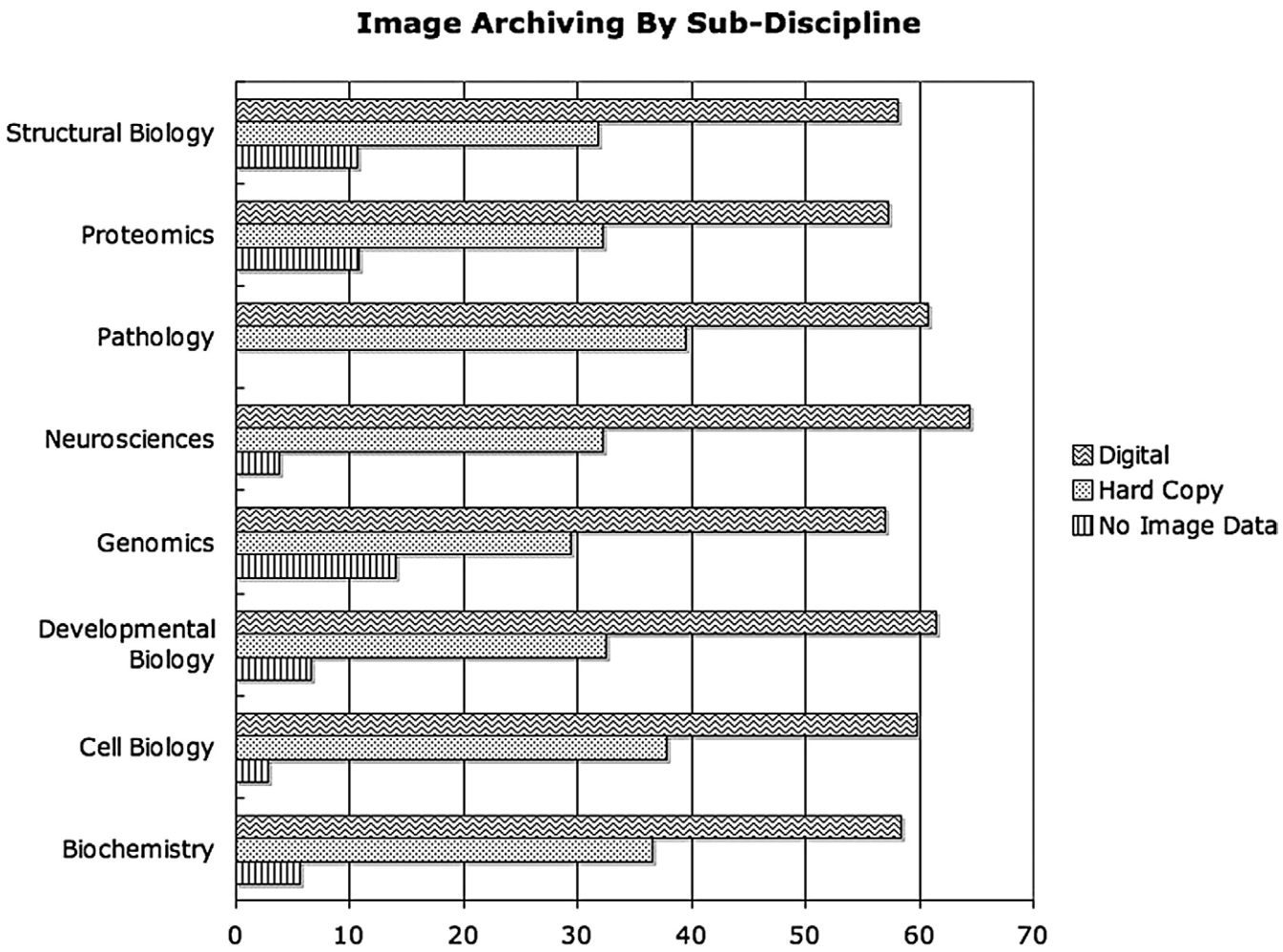

Figure 4. Image archiving by sub-discipline by percentage of respondents.

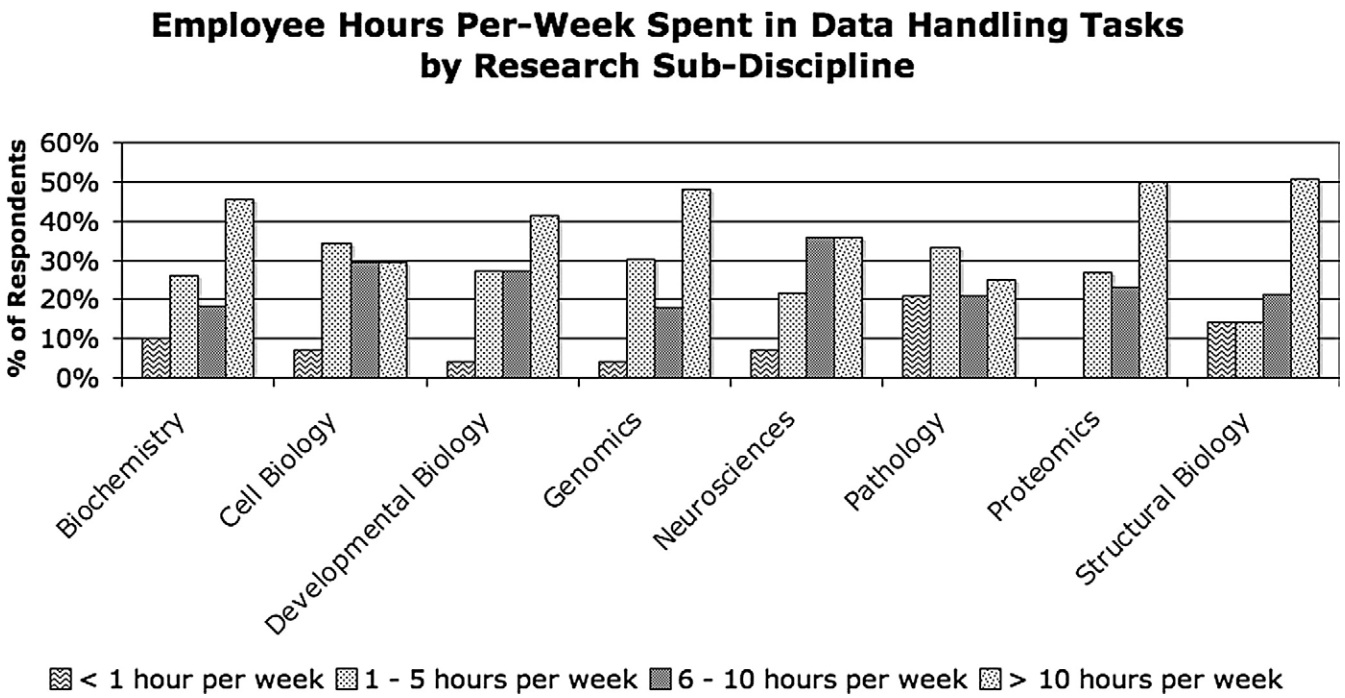

Figure 5. Employee hours spent per-week by sub-discipline and percentage of respondents.

> "Well, it's very cumbersome, I can't print anything, I'd have to paste it together. I end up just doing a freeze frame so that I can scroll this way."

Difficulty of Organizing, Storing, and Retrieving Data

A major component of biological research is the management of data. During the interviews, researchers frequently mentioned the complex and varied data formats they deal with throughout their research. We found it interesting that modern methods of information exchange (such as XML, MAGE) were not specifically mentioned, but we did not specifically ask for this information during the interviews. Researchers also discussed issues surrounding the increasing number of files, growing file sizes, multiple formats, indexing and annotation of data. A common theme surrounding data management and analysis was that many researchers preferred to utilize their own individual methods to organize data. The varied ways of managing data were accepted as functional for most present needs. Some researchers admitted to having no organizational methodology at all, while others used whatever method best suited their individual needs.

> "Yeah, I don't have them organized in any particular way and I don't have them linked to my databases, I just have them . . ."
> "They're not organized in any way—they're just thrown into files under different projects."
> "We have thousands of very large .tiff files which everyone organizes according to how they think they should be organized. We have kind of a lab standard like the filename begins with . . ."
> "I grab them when I need them, they're not organized in any decent way."
> "It's not even organized—a file on a central computer of protocols that we use, common lab protocols but those are just individual Word files within a folder so it's not searchable per se."

Although *ad hoc* organizational methods had been relatively successful to date, this lack of standardized methods for data organization was generally perceived as having a limited future. Specific technical problems associated with this lack of organization included: filenames being truncated when stored on central servers, the inability to easily store annotation data, and the lack of standardized formats or nomenclature for data sharing.

> "There are separate files but in independent computers that don't talk to each other."
> "Because what's happened is separate nodes have sprung up where you can have your own little partition of a drive somewhere and people will just password protect it with their name or whatever but technically everything is supposed to be shareable and viewable by everybody."
> "There have been times when folders have vanished or some piece of data was accidentally deleted or something gets corrupted and I think that that's why people are a little skeptical."
> "Right now, again, everybody has to do on their own and that it's always a pain to go and talk to 5 labs and go to each and say 'I want to get your datasets can you send me all of your original cell files and DAT files' and some of the labs won't do it."

As a result of relatively disorganized data, search and retrieval capability was also limited. Most researchers did not have a strategy for effective data retrieval. Many were not even aware of the capabilities of existing search and retrieval tools. Of those who were aware, few had implemented search capability.

> "One of the things that would be interesting for us to do is with PDF files—I have tons of PDF files saved on my desktop and they're not (collectively) searchable for the text that's in the PDF file."

Anticipated Data Management and Analysis Needs

In response to whether data handling and management currently caused a backlog in lab productivity (Question 28) 41% responded, "Yes", and 22% said "Somewhat." The problem was greater in larger labs where 51% responded, "Yes." Analyzed by specialty, data management issues impacting on productivity were most highly reported by structural biologists (62%), followed by proteomicists (59%) and genomicists (50%) (Fig. 5). The most frequently mentioned types of data causing management problems for survey respondents were microarrays (18%) and images (15%). Database capabilities (13%) and statistical analysis (13%) were also mentioned as obstacles to progress in the lab.

Eighty-two investigators responded to the open-ended survey question 35 asking them to detail their most urgent computational or data handling problem. Issues involving long-term digital archiving of a variety of data types were mentioned by more than 28% of respondents. Various forms of computerized data analysis (not involving a specific software product) were mentioned by 26%. Twenty-one percent of respondents felt that access to a specific software product would solve their most urgent problems while only 6% cited acquisition of hardware as a problem. Eighteen percent required access to some form of computer science or informatics expertise.

When asked to characterize how they located and evaluated tools to support these needs (Question 18), 41% stated they used the World Wide Web, 29% used a dedicated e-mail list with the remainder (24%) stating that this question was not applicable to their information management and analysis issues or that they used blogs or wiki's (combined 4%).

Analysis of the interview data identified two common anticipated needs: improved methods of managing large datasets, and improved ways to process and analyze data.

Need for Improved Methods of Managing Large Datasets

Some researchers were aware that their current *ad hoc* data management methods needed improvements. Several discussed the need for improvement of a whole laboratory information organization, moving away from organization based on individual preference and need to an established lab-wide data organization. A specific example of *ad hoc* organization was the common practice of researchers in the same lab creating custom spreadsheets without any common standard structure. The profusion of individually created spreadsheets containing overlapping and inconsistently updated data created a great deal of confusion within some labs. There was little consideration to future data exchange or submission requirements at the time of publication. Although researchers in the past have used spreadsheets containing a global data presentation to synthesize concepts and generate hypotheses, this approach became less feasible as data became more complex.

> "So we need really a way to store that in a database that is completely searchable, like you can search on any one of the items. So we try sometimes to put it in a filename but filenames become too long and when you store them to a server that doesn't like long filenames then they get truncated or they get misread. So this becomes, I think, one of our biggest problems."

> "For me it's mostly data organization and archiving—that's one issue, and then just analysis, how do you deal with all this data—it's still something that's very new to a lot of the labs, what do you do when you've got 10,000 data points for each time point - what's the best way to look at the data."

> "It makes me a little nervous the more databases that get generated . . . you know, which ones are really up to date . . ."

Need for Improved Methods of Analyzing and Processing Data

As mentioned earlier, though aware of the limitations of general-purpose applications such as spreadsheets, the majority of researchers continued to rely on them due to their ease of use, low cost, and familiarity. Researchers sought the convenience, low risk economics, and usability of these general-purpose tools while often recognizing that they were making a tradeoff against complex functionality that may be of use to them in the future.

> "I think anything that makes the interface more simple and straightforward is good."

> "I've sort of created my own little sad and pathetic database which is purely spreadsheet based, but it serves my needs. It's just that I am not an information architect in any way, or a database person, so I've sort of created it from the perspective of a scientist and having rows and columns and it's searchable and that's all I really need . . ."

> "But I would love to be able to export all this into something like [a spreadsheet] or some other program instead of me spending a week doing all this."

Similarly, many researchers relied on the general-purpose statistical analysis functions built into common spreadsheets despite there being better statistical analysis tools available in most domain areas.

Barriers to Addressing Data Management Issues

Three barriers to addressing data management and analysis emerged from the responses to question 35 in the survey and the personal perspectives provided during the interviews: financial burden of acquiring new expertise or tools, lack of time to invest in changing work practices to incorporate new technologies, and limited availability of institutional support.

Financial Burden of Acquiring New Expertise or Tools

Twenty-one percent of survey respondents felt that access to a specific software product would solve their most urgent problems while only 6% cited acquisition of hardware as a problem. Various forms of computerized data analysis (not involving a specific software product) were mentioned by 26% and 18% required access to some form of computer science or informatics expertise.

> "So the real problem is not so much the cost of the database and I don't remember what the seed price is but it's small. The real problem is supporting the cost of a database manager to support it."

Due perhaps to their lack of knowledge regarding the level of resources needed for handling complex data, many researchers underestimate the resources required for data handling in their grant proposals. Even if researchers had knowledge of technology, they commonly associated new technology with additional investment, either in terms of capital, training, or both. The single largest consensus on appropriate funding sources for tools was from indirect costs in research grants (37%)—this being the amount that

each awarded grant contributes to the parent institution for infrastructure such as research space, administration, and utilities—commonly 50% or more of direct costs. Eight percent felt that tools and support should be funded as subscriptions directly from research grants, 5% stated subscriptions funded from other sources, and a combined 38% thought that it would be appropriate to support this through all three of these categories (Question 20). Ten percent did not respond to this question, and 2% suggested other sources. In the interviews, limited funding made researchers wary of spending money on anything other than core research needs despite their awareness of the need for improved tools and/or additional support. Additionally, the lack of personal experience and/or the lack of success with previous tools contributed to a wariness of the value of investing in new solutions.

> "No, whenever I hear the word 'LIMS' I hear way too much money to deal with."
> "It's expensive, yes. So we have quite a history of attempts to use different relational databases—I've been here 7.5 years and we're going to be starting on the third one . . ."
> "I actually wanted the hospital to purchase it all and then I would just administer it. But since they fell down on that I was able to get 8 licenses—perpetual licenses—for DSGene, all the Wisconsin package, SeqWeb, everything else and one copy of all the databases for about 7 grand and then a server was 3, so for 10 I'm totally up and running and I'm not going to share them with anybody else because they haven't kicked in."
> "For whatever reason the university hasn't made generally available quicker access to these things and so it's extremely cumbersome and time consuming to do the kind of searches that we need to do since we haven't been able to find one of these big groups that have their own databases that we can hook up with."

Lack of Time to Invest in Changing Data Management Practices and Improving Training

In addition to limited financial resources, limited time also presented a significant barrier to improving data management and analysis processes. A common perception was that the time required for data management and training in the effective use of new technologies was not an integral part of experimental research. As a result, researchers discussed their frustration at having to spend increasing amounts of time to accomplish the physical management and analysis of data.

> "There are separate files but in independent computers that don't talk to each other. And the processing of the stuff is done through incredibly time wasteful methods on slow computers by untrained people that are doing it in small batches because that's all their computer can hold, or one at a time and then on pieces of paper, collating datasets"
> "And the whole data processing part of this was taking about 5 days. Not, you know, 40 hours, but from the point where the data was available and it had to go to this epidemiologist and then it had to come back, it generally took a whole week to get from the beginning to the end. And maybe it might have taken the individual people crunching the data maybe 5 hours or so of time, crunching."

The time commitment required for data transformations or changing workflows was perceived to be a significant lab expense, even when the financial cost wasn't necessarily an issue.

> "It's free in terms of money, it's not free in terms of your time."

These problems may have been exacerbated by the frequently reported high turnover of employees working in research labs.

Limited Availability of Institutionally Provided Expertise and Systems

More than half of the researchers speculated that improved support from their sponsoring institution would greatly improve the ability of individual researchers and their labs to focus on their research. Specific examples mentioned included lessening the financial burden for early stage researchers, greater access to and availability of institutionally supported data processing and analysis resource centers, and better technical support for both hardware and software development. Being within a large research university, there was a frequently held belief that the university could and should provide more basic support of data infrastructure than was perceived to be available. This was widely believed to be a potential way of relieving the limitations of time and budget discussed previously.

> ". . . It'd be sort of nice if the university had those tools that we didn't have to spend . . ."
> ". . . We should get something back from the university that helps us minimize our other costs if they could spend something and develop some resource for that."
> "I think the solution is the university needs to develop their own thing that everyone can use, because like I pointed out, the new people that come here, they've hardly any money and all of a sudden they have to shell out $5000 for all this software that they really need. If you kind of look at the way the whole system works, the university, it's advantageous to the university to provide new researchers with this stuff, because hopefully it will accelerate their research and bring in more indirects, right?"
> "They should have low cost shared storage—it's too expensive now—I think the University costs more than it does off campus—I don't know why storage needs to cost much of anything anymore."

Discussion

The themes and issues described in this work reflect a major shift in the way that information management and analysis has been traditionally conducted within the academic biomedical research laboratory. Prior to advances in high throughput technologies such as gene sequencing or microarray analysis, most researchers traditionally spent a great deal of time and effort focusing on the creation of highly specialized data. Today, however, individual investigators are increasingly required to study biological problems involving large amounts of diverse data that require special storage and analysis. Data management has become more complex with widely available high quality public scientific datasets and easier access to high throughput technologies at shared instrumentation locations. The increasing use of core service facilities within institutions to provide expertise such as biostatistics, or microarray assays—as well as mass-producing scientific technologies such as lab "kits"—has lowered many technical barriers, and has allowed investigators to generate and collect data outside their own discipline more easily. Yet despite these advances, the individual researcher or lab focused on specific problems

often lacks the time, funds, or experience to efficiently leverage these tools and services. At the present time biomedical research is in a phase where the quantity and the heterogeneity of data have exceeded many investigators' ability to analyze, or in some cases, even archive their own data.

These challenges have created opportunities for new research-support roles in biomedical research from the fields of informatics, statistics, and computer science. However, to date, the interaction between biomedical researchers, informaticians, and computer scientists has been marked by communication problems,[27,41,49–51] and there appears to remain an associated knowledge gap regarding what existing tools and resources are available as well as how to incorporate them into the research laboratory. From our interview and survey analysis, it is clear that from the individual researchers' perspective, there is inadequate institutional data management and analysis support for laboratory based biomedical researchers.

The reasons behind this lack of support appears to be a combination of social, technical, and fiscal factors that are perhaps in-part associated with the tradition of biomedical laboratory researchers being protective of their research. In general, most researchers we spoke with preferred to exert personal control over all aspects of data handling and organization. This reluctance to seek out and collaborate with a relational database expert or similar outside expertise may have led the researchers to overly rely on well-known but relatively generic spreadsheet applications. Overall, we found that the laboratory data management technology was bounded on the lower side by spreadsheets (extremely flexible but limited in capacity and capability) and on the upper side by relational databases (high in capacity and capability but inflexible). However, virtually every investigator who was not familiar with relational databases preferred to use a spreadsheet if it allowed them to easily manipulate and manage their data—despite a lack of data analysis and scaling that would be available to them by using database tools. Though there is significant research to address this gap by providing applications[52–54] that integrate ease of configurability with powerful querying and presentation capabilities, it remains to be seen if they will be adopted by typical investigators. One evolving example is the open source analysis environment of Bioconductor,[55] which provides broad biological analysis and data integration capabilities, and has considerable general support in the research community facilitated through researcher contributed plug-ins that address specific analysis tasks. As of this writing, however, Bioconductor is a sophisticated development environment with a significant learning curve that may be beyond the programming capabilities of many researchers.

Collaboration between laboratories may encourage sharing of more sophisticated data management and analysis strategies; however, collaboration is difficult due to the highly focused and non-overlapping nature of different study groups, where each group has its own terminologies and nuances.[28–30,56] Our results indicate that the needs of individual investigators have a great deal of overlap, but the needs of different labs and sometimes even within labs vary widely. Several researchers were convinced that "everybody needs the same things" but their efforts to use software solutions used by similar labs were not successful. A common complaint was that those tools "only work for that lab."

It is possible that the researchers' perceived shortcomings of the institutional support facilities were in fact due to the unique characteristics of individual lab needs. Many of the researchers who did attempt to utilize institutional resources had unsatisfactory experiences and eventually decided these resources could not be utilized in their own study. Based on our analysis, it appears that increased availability of centralized expertise and resources might not be sufficient to address the diverse needs of labs when coupled with their desire to have flexible and customizable solutions. Tools specific to particular analysis tasks are difficult to centralize and provide without particular domain knowledge. Unique solutions for each research area should be studied individually without the added requirement of generalization; however, solutions that are unique to individual domain areas should be able to utilize a common information support infrastructure as yet to be formed. Establishment of such an infrastructure could facilitate the convergence of commonly used resources and allow for increased collaboration and information sharing among researchers, while still supporting individually unique research goals.

Many investigators felt that some form of data management systems should be provided more centrally either by their sponsoring institutions or by the funding agencies. Usually lacking these services, researchers tended to avoid the time and costs that are perceived to be associated with implementing more sophisticated systems. In interviews many researchers stated that in return for indirect costs from their grants, the institutions provided much less data support than they wanted. The researchers also expressed the view that if either funding agencies or institutions provided more support, they would be able to save badly needed research funds by preventing duplication of effort and expenditure for analytic software and computer science expertise. If the trend of increasing size and complexity of datasets continue—which is probable—funding entities may call for increased institutional support of data management and analysis as they do for lab space.

Conclusion

The aims of this research were to evaluate the data management and analysis needs of biomedical researchers and identify barriers to addressing those needs. We also sought to better understand what information management needs could be generalized, if any. While we recognize that the methods used to gather this data reached only a small percentage of the researchers actively involved in this single research location, these individuals had self-selected themselves to be on this e-mail list out of a shared interest in on-going biomedical data management, analysis, and education issues at this institution. We believe that the themes we identified reflect and characterize the common perceptions of many researchers encountering on-going data analysis problems and difficulties working with large datasets. Despite its limitations, this study has provided a basis for further research at our institution to identify likely solutions to these data management and analysis problems.

In summary, we suggest that at least some data management needs of biomedical researchers will best be served by

improving and providing basic level tools such that scientists can solve their own problems, such as designing specialized applications as plug-ins to evolving commonly used frameworks such as Bioconductor. We also suggest that both institutions and informaticians interested in supporting laboratory information management needs should increase focus on three components: 1) facilitate and encourage the use of modern data exchange models and standards, which will allow individual researchers with specific problems to have a common layer of interoperability and analysis; 2) improve the ability of researchers to maintain provenance of research data and models as they evolve over time; 3) develop and support biomedical information management service cores that can facilitate both of these previous components while providing a spectrum of support options that could be used to address the range of unique individual researcher needs. These basic infrastructure components could provide considerable secondary benefits, such as increased collaboration and greater leveraging of existing research personnel for core science roles.

References ■

1. NIH. NIH Roadmap on Translational Research. 2005. Available at: http://nihroadmap.nih.gov/clinicalresearch/overview-translational.asp. Accessed February 22, 2006.
2. NIH. NIH RFP for Institutional Clinical and Translational Science award. 2005. Available at: http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-06-002.html. Accessed: March 14, 2006.
3. NIH. Biomedical Information Science and Technology Initiative (BISTI). 2001. Available at: http://www.bisti.nih.gov/. Accessed: March 14, 2007.
4. NSF. Science and Engineering Information Integration and Informatics. 2004. Available at: http://www.nsf.gov/pubs/2004/nsf04528/nsf04528.htm. Accessed: March 14, 2007.
5. NIH Announces Draft Statement on Sharing Research Data. 2002. Available at: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-035.html. Accessed: March 14, 2007.
6. Cadeg E, Louie B, Myler P, Tarczy-Hornoch P. BioMediator Data Integration and Inference for Function Annotation of Anonymous Sequences. Pacific Symposium on Biocomputing. Maui, Hawaii, 2007.
7. Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P. Data integration and genomic medicine. J Biomed Inform. 2007;40:5–16.
8. Mei H, Tarczy-Hornoch P, Mork P, Rossini AJ, Shaker R, Donelson L. Expression array annotation using the BioMediator biological data integration system and the BioConductor analytic platform. AMIA Annu Symp Proc. 2003:445–9.
9. Donelson L, Tarczy-Hornoch P, Mork P, et al. The BioMediator system as a data integration tool to answer diverse biologic queries. Medinfo. 2004;11(Pt 2):768–72.
10. Jakobovits RM, Rosse C, Brinkley JF. WIRM: an Open Source Toolkit for Building Biomedical Web Applications. J Am Med Inform Soc. 2002;9(6):557–70.
11. Li H, Gennari JH, Brinkley JF. Model Driven Laboratory Information Management Systems. American Medical Informatics Association. Washington DC, 2006.
12. Fong C, Brinkley, J. Customizable Electronic Laboratory Online (CELO): A web-based data management system builder for biomedical laboratories. Fall Symposium of the American Medical Informatics Association. Washington DC, 2006.
13. Oinn T, Addis M, Ferris J, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004;20(17):3045–54.
14. Jeng S, Wang K, Barbero J, Brinkley J, Tarczy-Hornoch P. A Pilot Bridging Data Integration and Analytics: BioMediator and R. AMIA Annual Symposium. Washington DC, 2005.
15. Jones C. Patterns of Software System Failure and Success. Stamford, CT: International Thompson Computer Press, 1996.
16. Brooks FP. The Mythical Man-Month: Essays on Software Engineering. Boston, MA: Addison-Wesley Professional, 1995.
17. Seaman C. Communication and Organization in Software Development: An Empirical Study. IBM Systems Journal. 1997;36:550–63.
18. Seaman C. Qualitative Methods in Empirical Studies of Software Engineering. IEEE Transactions on Software Engineering. 1999;25(4):557–72.
19. Gittens R, Hope, S, Williams, I. Qualitative Studies of XP in a Medium Sized Business. Proceedings of the 2nd Conference on eXtreme Programming and Flexible Processes in Software Engineering. Cagliari, Italy, 2001.
20. Lindgaard G, Dillon R, Trbovich P, et al. User needs analysis and requirements engineering: theory and practice. Interact Comp 2006;18(1):47–70.
21. Bryman A. Integrating quantitative and qualitative research: how is it done? Qual Res 2006;6:97–113.
22. Boverhof DR, Zacharewski TR. Toxicogenomics in risk assessment: applications and needs. Toxicol Sci. 2005;89:352–60.
23. Korjonen-Close H. The information needs and behaviour of clinical researchers: a user-needs analysis. Health Info Libr J 2005;22(2):96–106.
24. Strasberg HR, Tudiver F, Geiger G, Keshavjee KK, Troyan S. Moving towards an electronic patient record: a survey to assess the needs of community family physicians. AMIA Annu Symp Proc 1998:965–9.
25. Tanner C, Eckstrom E, Desai SS, Ririe MR, Bowen JL. Uncovering frustrations. a qualitative needs assessment of academic general internists as geriatric care providers and teachers. J Gen Intern Med. 2006;21(1):51–5.
26. Rosenal TW, Forsythe DE, Musen MA, Seiver A. Support for information management in critical care: a new approach to identify needs. Proc Annu Symp Comput Appl Med Care. 1995:2–6.
27. Forsythe DE. Using ethnography to investigate life scientists' information needs. Bull Med Libr Assoc. 1998;86(3):402–9.
28. Ammenwerth E, Shaw NT. Bad health informatics can kill--is evaluation the answer? Methods Inf Med. 2005;44(1):1–3.
29. Kaplan B, Shaw NT. Future directions in evaluation research: people, organizational, and social issues. Methods Inf Med. 2004;43(3):215–31.
30. Kaplan B. Evaluating informatics applications - some alternative approaches: theory, social interactionism, and call for methodological pluralism. Int J Med Inform. 2001;64:39–56.
31. Yarfitz S, Ketchell DS. A library-based bioinformatics services program. Bull Med Libr Assoc. 2000;88(1):36–48.
32. Tran D, Dubay C, Gorman P, Hersh W. Applying task analysis to describe and facilitate bioinformatics tasks. Medinfo 2004;11(Pt 2):818–22.
33. Anderson N, Ash J, Tarczy-Hornoch P. A qualitative study of the implementation of a bioinformatics tool in a biological research laboratory. Int J Biomed Inform. 2006 (e-pub ahead of print).
34. Bartlett J, Toms E. Developing a protocol for bioinformatics analysis: an integrated information behaviour and task analysis approach. J Am Soc Inform Sci Technol 2005;56(5):469–82.
35. Bartlett J, Toms, E. Discovering and structuring information flow among bioinformatics resources. Proc 26th Annu Intern ACM SIGIR. Toronto, Canada, 2003.

36. Ash JS, Sittig DF, Seshadri V, Dykstra RH, Carpenter JD, Stavri PZ. Adding insight: a qualitative cross-site study of physician order entry. Int J Med Inform. 2005;74(7–8):623–8.
37. Ash JS, Fournier L, Stavri PZ, Dykstra R. Principles for a successful computerized physician order entry implementation. AMIA Annu Symp Proc. 2003:36–40.
38. Crabtree B, Miller W. Doing Qualitative Research. 2nd ed: Thousand Oaks, CA: Sage Publications, 1999.
39. Fisher K, Erdelez S, McKenchie L. Theories of Information Behavior. Medford, NJ: Information Today, 2005.
40. Wolcott HF. Writing Up Qualitative Research: Thousand Oaks, CA: Sage Publications, 2001.
41. Ash JS, Stavri PZ, Kuperman GJ. A consensus statement on considerations for a successful CPOE implementation. J Am Med Inform Assoc. 2003;10(3):229–34.
42. Wooldridge A. The Brains Business. The Economist. September 8, 2005.
43. University of Washington Office of Research. 2006. Available at: http://www.washington.edu/research/about.html. Accessed October 23, 2006.
44. WebQ Home Page. Available at: http://catalyst.washington.edu/tools/web_q.html. Accessed February 1, 2006.
45. Catalyst Home Page. Available at: http://catalyst.washington.edu/. Accessed February 1, 2006.
46. University of Washington BioResearcher Toolkit. Available at: http://healthlinks.washington.edu/index.jsp?id=210BCCB7-511A-4C6B-8B40-DFC47AABEA7F. Accessed February 1, 2006.
47. Miles M, Huberman AM. Qualitative Data Analysis: An Expanded Sourcebook. Thousand Oaks, CA: Sage, 1994.
48. Jakobovits R, Soderland SG, Taira RK, Brinkley JF. Requirements of a Web-based experiment management system. Proc AMIA Symp. 2000:374–8.
49. Arnstein L, Grimm R, Hung C, et al. Systems Support for Ubiquitous Computing: A Case Study of Two Implementations of LabScape. Proc First Intern Conf Perv Comp. Germany: Springer-Verlag, 2002.
50. Flowers S. Software Failure: Management Failure: Amazing Stories and Cautionary Tales. Hoboken, NJ: Wiley and Sons, 1996.
51. Forsythe DE. Using ethnography to build a working system: rethinking basic design assumptions. Proc Annu Symp Comput Appl Med Care. 1992:505–9.
52. Pittendrigh S, Jacobs G. NeuroSys: a semistructured laboratory database. Neuroinform. 2003;1(2):167–76.
53. Marenco L, Tosches N, Crasto C, Shepherd G, Miller PL, Nadkarni PM. Achieving evolvable Web-database bioscience applications using the EAV/CR framework: recent advances. J Am Med Inform Assoc. 2003;10(5):444–53.
54. Li H, Brinkley J, Gennari J. Semi-automatic Database Design for Neuroscience Experiment Management Systems. Proceedings of the MedInfo 2004 Conference, San Francisco, CA, September 7–11, 2004.
55. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.
56. Forsythe DE. New bottles, old wine: hidden cultural assumptions in a computerized explanation system for migraine sufferers. Med Anthropol Q. 1996;10(4):551–74.