MU Libraries
Life Sciences Research Task Force
Report on Activities

November 2007

Prepared by:
Kate Anderson, MLS,
MaryEllen Cullinan Sievert, PhD,
Brenda J. Graves-Blevins, MLS,
Deborah H. Ward, MLS

Contents of Report

I. EXECUTIVE SUMMARY

In January 2006, the Director of the MU Libraries charged the Life Sciences Research Task Force with creating a systematic approach to information service delivery for persons engaged in life sciences research on the MU campus, beginning with personnel working at the Christopher S. Bond Life Sciences Center (Bond LSC). The Task Force focused on exploring the scientists' needs and creating outreach services tailored to their needs with the ultimate goal of providing enhanced service to all life science researchers across campus.

Over an 18-month period, we examined programs at other institutions, conducted a series of focus groups with a total of 40 people in the Bond LSC and consulted with a Molecular Biosciences Information Specialist on ways in which the MU Libraries could support bioinformatics work on campus.

Conclusions:

- Focusing on life sciences research brought together librarians from across campus, encouraging collaboration in new ways.
- Researchers often do not know what resources the library offers and what services the librarians can provide.
- Bioinformatics support is qualitatively different than the type of support currently available through the MU Libraries.
- Work done by the Life Sciences Research Task Force has laid the groundwork for future partnerships and collaborations with researchers at the Bond LSC.

Recommendations:

- Designate a librarian to be the contact point for all those in the Bond LSC.
- Continue to provide new or enhanced services based on user needs.
- Continue to pursue how MU Libraries can support bioinformatics research, specifically via the creation of a Molecular Biology Specialist position.

II. BACKGROUND AND CHARGE OF TASK FORCE

In the late 1990s, the University of Missouri (MU) created a strategic plan that would use funds from the Missouri Legislature's Mission Enhancement program and other sources to "improve the life sciences, including construction of necessary instructional and research facilities, the addition of new faculty, and other enhancements" (SPRAC report, 1999). Concurrently, there was a call for more interdisciplinary research on campus. The interdisciplinary Christopher S. Bond Life Sciences Center (Bond LSC), completed in 2005, became the physical manifestation of these strategic directions.

For the MU Libraries, a building like the Bond LSC represents both a challenge and an opportunity. Traditionally, different departments have been served by different campus libraries. However, with Principal Investigators from a multitude of departments and colleges now in one building, the Libraries needed to reassess common services. Also, as life sciences research becomes ever more prominent on the MU campus, the MU Libraries needed a mechanism to systematically investigate the needs of this type of researcher. For example, the workflow of a life sciences researcher can be very complex as there are a vast array of information resources which contain information pertinent to a specific project (see Appendix A). Most scientific researchers do not have the time to become aware of all the available resources, and they should be able to rely on information experts to direct them to the most appropriate resources for their research.

In January 2006, the MU Libraries created the Life Sciences Research Task Force with the charge of examining the needs of life sciences researchers and developing and adjusting library services accordingly. Because of the great number of life sciences researchers on campus, the Task Force focused its attention on the Bond LSC. This group spent approximately eighteen months meeting about these issues. See Appendix B for list of the members of the Task Force.

III. TASK FORCE ACTIVITIES

The members of the Life Sciences Center Research Task Force decided on a multi-pronged approach to investigating the needs of life sciences researchers on campus that included: education, focus groups, and expert consultation.

Approach One: Education
Task Force members educated themselves on the research happening in the Bond LSC and on what services other libraries were offering. Educational activities included:

- Reviewing the literature on library services to life sciences researchers
- Investigating programs at other libraries across the country
- Touring the LSC as a group

- Traveling to Becker Library at the Washington University School of Medicine to meet with the director and the two scientists hired by the library as bioinformatics specialists.

All of these activities informed our thinking as we prepared to meet with personnel at the Bond LSC.

Approach Two: Focus Groups
After obtaining IRB approval, two members of the Task Force conducted five focus-group interviews. Information gathered from the focus groups was presented as a poster at the 2007 Annual Meeting of the Medical Library Association in Philadelphia (see Appendix C). See Section IV. for detailed information on the focus groups.

Approach Three: Expert Consultation
On July 30 and 31, 2007, Diane Rein, Assistant Professor of Library Science and Assistant Life Sciences Librarian at Purdue University, visited MU Libraries to conduct a six-hour introductory workshop on bioinformatics and to participate in a discussion on how MU Libraries can deliver bioinformatics services to life sciences researchers on campus. Dr. Rein's visit was supported by funding from the Institute of Museum and Library Services through a "Bring in the Expert Grant" administered by the Missouri State Library. See Appendix D for a report of that visit.

IV. EXPLORING THE INFORMATION NEEDS OF LIFE SCIENCE RESEARCHERS AT THE CHRISTOPHER S. BOND LIFE SCIENCES CENTER

Due to the diverse nature of research in the Bond LSC, focus groups were an appropriate way to make new connections to these researchers. Additionally, meeting with researchers in their building let them know that the Libraries were interested in creating new kinds of partnerships.

Focus groups generally give the participants the opportunity to create the discussion. Therefore, having only a few general questions keeps the discussion going. For these interviews, we used only three questions:

1. What is your information environment like?
2. How do you manage the information you need in your research?
3. What are the gaps in your information environment?

Two members of the Life Sciences Research Task Force conducted five focus groups, with a total of 40 participants from the Bond LSC. The composition of three of the focus groups were designed to give a fairly broad view of information needs of the researchers. The last two groups were to give an in-depth look at a group working on similar projects under the supervision of a single faculty member.

Composition of the Focus Groups:

1. Graduate Students and Post-doctoral Fellows: **8 participants**

2. Junior Faculty: **10 participants**

3. Bond LSC Administrators: **5 participants**

4. Individual Lab: **10 participants**

5. Individual Lab: **7 participants**

The interviewers alerted the interviewees that there would be time at the end of the session for them to request specific new journals and that their saying they never used the library would be acceptable. The interviewers realized during the second interview that some themes were going to recur among all groups being interviewed.

General conclusions from all focus groups:

- The researchers at the Bond LSC did not know all of the resources the library offers and what services the librarians can provide.

    - For example, several doctoral students told the interviewers that the journal *Cell* was not available electronically in full text on campus. *Cell* is available electronically.

    - The faculty interviewed did not know that they could make arrangements for their graduate students to use the Copy Center at Ellis Library or check out books for them.

    - Few knew that the library provided instruction in specific resources or software and that they could request such training for their laboratories be provided at the Bond LSC rather than in the library.

    Immediate Follow-Up: Based on information from these focus groups, the librarian interviewer offered classes in EndNote to three laboratories and answered any questions that arose and could not be answered immediately.

- Initially the researchers did not think having a single contact person at the library would be useful. However, they changed their minds as they thought about the current complex structure. For example, those with departmental affiliation served by Ellis Library could go to different librarians depending on whether their main appointment was in the College of Arts & Sciences or the College of Agriculture, Food and Natural Resources.

    Recommendation: Based on this feedback, one of our recommendations is to designate a librarian, or a primary librarian with an alternate, as the point of contact for all life sciences researchers in the Bond LSC.

- During the interviews it became clear that the research in the laboratories at the Bond LSC was narrow and deep. As a result there would always be needs specific to any one laboratory or group of researchers. One size definitely would not fit all.

Responses to individual questions
Question 1: What is Your Information Environment Like?

There was consensus about certain needs and considerable diversity on others. The individual laboratories showed more consistency than the broader groups as would be expected.

- The most obvious theme across all interviews was the use of and need for the full-text of journal articles.
    - In addition, some wanted color in the illustrations if it was available in the printed version.
    - Most wanted not only the current issues, but also back issues.
    - Several noted that when they requested interlibrary loan they did not get the articles in an electronic format useful to them, i. e. they received them as photocopies rather than as PDF files.
- For some their needs were best answered by Google rather than more "scientific" sources, in part because the information needed was so scattered. Almost every one of those interviewed used Google at times because it was so easy, so familiar. One person remarked that without Google he would have to go to multiple databases (at least 3) to cover all the places where relevant information might be indexed and that "would drive [him] crazy."
- Most of those interviewed felt that they did not need instruction on accessing information from specialized sources. However, the literature about services to similar patrons at other institutions always mentioned how well-received the instruction was and how the classes offered generally filled up before the deadline.

  Immediate Follow-Up: When the Libraries and the Bond LSC decided to jointly sponsor training from the National Center for Biotechnology Information (NCBI) of the National Library of Medicine, many were enthusiastic.
- Aside from Google, there was not a single source to which all went for information.
    - Some used PubMed; some used Web of Knowledge, and a few used an alerting system of some kind. Several used the databases offered by NCBI.
    - Often one or two in a group would use other biomedical sources, including MolBio.net specific to molecular biologists, while others would use only very

narrowly focused web resources of interest only to a small group of researchers worldwide.

Question 2: How Do You Manage Your Information?

The results for this question brought considerably more agreement.

- For recording what was happening in the laboratory, everyone used the traditional laboratory notebook.
    - One person said he had experimented with an electronic notebook and was using both currently. Another said he wrote his lab notes with Word and then pasted them into his lab notebook.
    - One faculty member would be interested in a master notebook in which everything stored in individual notebooks was also stored in the master notebook. He knew of no software which did this.
    - Another remarked that while scientists may receive much of their information in electronic form, they still store it on paper.
    - Along the same lines, several reported printing pieces of information, particularly charts and figures, and then gluing or taping the printout into the notebook.
- For keeping track of bibliographic references, everyone used EndNote.
    - However, there was a wide disparity between the abilities and comfort of those using it.
    - Some (particularly faculty) were proficient; others were beginners.

Question 3: What Are the Gaps in Your Information Environment?

For this question, responses that appeared diverse at first actually fell into broad groups.

- Several people mentioned the need for experts in the scientific software used in their laboratories. What seems to be the broader issue is that the Bond LSC needs people who are experts in several kinds of laboratory-support software, e.g. a database for recording changes in animal data over time. An information expert who had been trained in the selection and acquisition of information resources could create a database of what is used and be able to provide guidance in the use of currently available software and in the selection of new software which might be useful to more than one laboratory. The bioinformatics librarians at Washington University have done the latter, and their efforts have been well received by research personnel.

- Also, researchers mentioned the need for software to assist with specific scientific tasks. Again, a grouping of these needs might provide a basis for the selection of some generalized scientific software packages which could then be configured for the needs of individual laboratories.

- The fact that the scientists at the Bond LSC do not know what software packages are being used in different laboratories or that graduate students in one laboratory did not know what software support package others in the laboratory were using suggests that increased communication both between and within the laboratories would be helpful to all.

V. OVERALL CONCLUSIONS AND RECOMMENDATIONS

Interviews with the five focus groups representing both individual laboratories and cross-laboratory subsets of researchers in the Bond LSC revealed a number of current information needs.

Services Possible with Existing Personnel

- The MU Libraries can designate a librarian—or a primary contact with an alternate—to serve as an Information Specialist for the Bond LSC.

    - This librarian should be available in the Bond LSC itself, as needed, since offering the services from the libraries will not be sufficient. Most of those interviewed did not want to go to the library itself.

    - This designated librarian would most appropriately be the librarian involved in the focus groups because of her understanding of science, scholarly communication and library resources. She also has knowledge and familiarity with online resources so that she will be comfortable working in this environment and will be able to interact with the computer people in the Bond LSC.

    - Instructional services for the research personnel, offered to both individuals and to groups, should be ongoing and tailored to individual needs. Also, it might be advisable to regularly (once or twice per year) send out messages about library services to people new to the Bond LSC.

    - If the librarian working with the Bond LSC personnel is to be successful, she must be in close communication with the faculty, the doctoral students, and post-doctoral fellows.

    - Importantly, the designated librarian cannot simply add these new responsibilities to her workload but must be relieved of some of her current duties so that she can serve the researchers adequately.

Services Possible with Additional Personnel

What we learned from the focus groups, from our reading, and from our consultations with bioinformatics or molecular biology specialists is that the MU Libraries could offer improved services with a different kind of librarian, a Molecular Biology Specialist focused on bioinformatics. See Appendix E for a draft of a position description.

For the Bond LSC, this person could provide new (and sometimes better or more appropriate) information services. The Molecular Biology Specialist could help researchers as they produce papers and grants in finding all the appropriate resources. She/he could also help in finding and teaching, new scientific software which may help the research in the laboratory. The literature suggests a number of other roles this person could assume for the benefit of the Bond LSC researchers, such as teaching a class on information resources appropriate for specific kinds of researchers.

In the long run, the creation of a position of a Molecular Biology Specialist to serve the researchers—at the Bond LSC initially and with expansion to life sciences researchers across campus—will have valuable outcomes for both the MU Libraries and the MU campus. For the Libraries this person, along with some ongoing work at the J Otto Lottes Health Sciences Library, will provide the models for the libraries to move their services out to their users. Also, housing a Molecular Biology Specialist within the MU Libraries can, we feel, provide a neutral resource for bioinformatics issues across campus.

Appendix A

Excerpts from article about workflow of scientists: Patrick TB, Craven CK, Folk LC. The need for a multidisciplinary team approach to life science workflows. JOURNAL OF THE MEDICAL LIBRARY ASSOCIATION 95 (3): 274-278 JUL 2007.

It is no more reasonable to expect biologists to be experts in the metadata of biological information resources than it is to expect librarians to be experts in biology. Thus, because even simple, apparently similar information retrieval workflows may produce different results, a multidisciplinary team approach to authoring, vetting, and using life science workflows is needed. Such teams must include experts in the primary science and experts in the metadata characterizing the information resources.

The importance of librarians as metadata experts in life science research was recognized by the Human Genome Project in 1997 [23]. Unfortunately, almost a decade later, the library remains largely excluded from the mainstream of life science research: very few universities offer bioinformatics end-user support services through the library [24]; demand is generally not great for such services when offered [25]; and molecular biology students in particular do not choose the library as their preferred source of information about bioinformatics databases [26].

The life science information space is growing extremely rapidly, largely facilitated by "the breakdown of the traditional barriers between academic disciplines and the application of technologies across these disciplines" [27]. Similarly, breaking down the barriers between "scientist" and "librarian" and fostering the interdisciplinary and synergistic combination of their respective expertise in the development and use of life science workflows are crucial to achieving full and optimal exploitation of the life science information space.

Complete article available below and at:
http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=1924931

Biomedical Library, and Adjunct Instructor, Department of Biomedical Informatics; **Nunzia Bettinsoli Giuse, MD, MLS, AHIP, FMLA,** nunzia.giuse@vanderbilt.edu, Director, Eskind Biomedical Library, and Professor, Departments of Biomedical Informatics and Medicine; Vanderbilt University Medical Center, Nashville, TN 37232

Received November 2006; accepted January 2007

The need for a multidisciplinary team approach to life science workflows*†

Timothy B. Patrick, PhD; Catherine K. Craven, MLS, MA; Lillian C. Folk, MS

See end of article for authors' affiliations.

DOI: 10.3163/1536-5050.95.3.274

INTRODUCTION

Information retrieval for life science research (a broad rubric encompassing many traditional disciplines such as biochemistry, botany, cell biology, and molecular biology [1]) often involves the use of combinations of multiple information resources. Such combinations have been called "workflows" [2, 3] and may include factual databases such as Genbank [4], literature databases such as Entrez-PubMed [5], and analysis tools such as the Basic Local Alignment Search Tool (BLAST) [6]. Information resources can be combined in different ways toward the same goal; varying combinations may produce different results for the same research question. Combinations that produce different results may appear equivalent to a scientifically sophisticated user who lacks knowledge of metadata about the resources that may indicate the possibility of varying results. In addition, a user who pursues only a single combination of resources may not even realize that another combination might produce different results.

This study's objective was to compare the results of three intuitively plausible and seemingly similar workflows for retrieving gene function information, with the goal of illustrating the importance of library science in bioinformatics and the need for a multidisciplinary team approach to authoring, vetting, and using life science workflows.

* Based on a presentation at MLA '06, the 106th Annual Meeting of the Medical Library Association, Phoenix, AZ, May 19–24, 2006.
† This research was supported by a Medical Library Association Donald A. B. Lindberg Research fellowship and the National Library of Medicine Biomedical and Health Informatics Research Training grant 2-T15-LM07089-11.

Supplemental Figures 1, 2, 3, and 4 are available with the online version of this journal.

METHODS

Microarray analysis is a high-throughput experimental technique that engenders significant information retrieval requirements [7]. One use of microarrays is analyzing gene expression: raw data from the microarray are statistically analyzed to determine which genes show significant changes in expression, with one or more lists of genes as the final result. Interpreting the biological meaning of this result often necessitates retrieving information from other sources about the function of the listed genes. Microarray analysis, therefore, is one example of a domain in which information from the biological literature must be integrated with information contained in sequence and other databases.

For some microarray analyses, each gene has a related representative DNA sequence. The identifier of that DNA sequence (its nucleotide sequence accession number, hereafter, "accession number") may be used to search for information about the function of the associated gene. This study compared three workflows that used accession numbers as starting points and utilized linkages among PubMed and other Entrez databases [8]. Although using accession numbers to search for gene function information has problems [9], the workflows compared here have been selected as simple, intuitively plausible strategies similar to some of those the authors have seen used in practice. Other workflows, using other starting points or information resources, are also possible and potentially useful.

This study used a list of 251 accession numbers representing genes determined to be of interest in a microarray experiment related to muscle recovery after immobilization (NIH grant AG18881) [10–12]. The genes on the list represented an example of real-world microarray results for which researchers might need to retrieve gene function information. The list of accession numbers was used as the test-set against which workflows were executed and their results compared.

Description of the three workflows

The three workflows are depicted in Figure 1 (available online). Each starts with an accession number (e.g., *M29293*), denoted as "*xxxxxxx*."

Workflow 1: PubMed only. The Entrez-PubMed "Secondary Source" or SI field (which identifies secondary data sources and associated accession numbers discussed in MEDLINE articles) [13] was searched using a query of the form *genbank/xxxxxxx[si]*. The result was a set of PubMed records, represented here as a set of PubMed IDs (PMIDs). For example, the query "*genbank/M29293[si]*" retrieved PMID *2532363*.

Workflow 2: Nucleotide-PubMed. Entrez-Nucleotide [14] was searched using a query of the form "*xxxxxxx*" and retrieved nucleotide sequence records that might provide links to other resources. Two types of links, *PubMed links* and *PubMed Central links*, were pursued. *PubMed links* led to Entrez-PubMed and a set of

Table 1
Workflow results and comparisons

	Comparison 1: nucleotide accession numbers associated with one or more PubMed IDs	Comparison 2: PubMed IDs retrieved	Comparison 3: nucleotide sequence accession number–PubMed ID pairs retrieved
Number retrieved by:			
Workflow 1	49	72	73
Workflow 2	126	101	192
Workflow 3	45	267	301
Total retrieved	127	338	464
Number retrieved by workflows:			
1 only	0	18	18
2 only	72	38	129
3 only	1	219	254
1 AND 2 (but NOT 3)	10	15	16
1 AND 3 (but NOT 2)	0	0	0
2 AND 3 (but NOT 1)	5	9	8
1 AND 2 AND 3	39	39	39

PMIDs. *PubMed Central links* led to PubMed Central (the Entrez full-text repository) [15] and a set of PubMed Central records. These records had a *PubMed links* option, which provided a set of PMIDs corresponding to the PubMed Central records. For example, the query "*M29293*" led, via the *PubMed links*, to PMID *2532363* and via the *PubMed Central links*, to PMIDs *15644144* and *2532363*.

Workflow 3: Gene-PubMed. Entrez-Gene [16] was searched using a query of the form "*xxxxxxx[NACC]*" ([*NACC*] was used to unambiguously declare *xxxxxxx* an accession number). The result was the record for a gene that might provide links to other resources. As before, only *PubMed links* and *PubMed Central links* were pursued. For example, the query "*M29293 [NACC]*" retrieved an entry for the gene *Snrpn*. That gene entry included both *PubMed links* and *PubMed Central links*. In this example, both the *PubMed links* and the *PubMed Central links* led to PMIDs *12477932* and *2532363*.

Workflow comparison procedures

Previously, the 251 accession numbers were searched using Java implementations of the 3 workflows, and results were partially reported [17]. Between July 14 and 24, 2006, the search results were manually verified and updated. For each workflow, the PMIDs retrieved by each accession number were recorded. For workflows 2 and 3, whether the PMIDs could be retrieved via the *PubMed links* or *PubMed Central links* was also recorded.

Three aspects of the workflows were compared: which and how many accession numbers successfully retrieved one or more PMIDs, which and how many PMIDs were retrieved, and which and how many unique pairings between a particular accession number and a particular PMID (hereafter, "accession number–PMID pairings") were produced. The overall output of each of the three workflows was compared. In addition, for workflows 2 and 3, the results of following the *PubMed links* and *PubMed Central links* paths were compared. Because workflow 1 involved direct search of PubMed, this workflow had no alternative paths to the literature.

Statistical analysis

Agreement between pairs of workflows was assessed using Cohen's kappa [18] (denoted K). Statistical calculations were performed using SPSS [19]. The P value for each individual comparison was multiplied by nine to adjust for multiple comparisons [20]; adjusted P values < 0.05 were considered significant. Significant comparisons were interpreted as suggested by Byrt [18].

RESULTS

Tables 1, 2, and 3 present the aggregate study results. Figures 2, 3, and 4 present the results of comparisons 1, 2, and 3, respectively.

Comparison 1: Which and how many accession numbers were successfully used to retrieve one or more PubMed IDs (PMIDs) using the different workflows?

Overall results. PMIDs were associated with 127 accession numbers: 49 by workflow 1, 126 by workflow 2, and 45 by workflow 3. In terms of overlap, 39 accession numbers were associated with PMIDs by all 3 workflows.

***PubMed links* and *PubMed Central links* paths.** In workflow 2, 83 accession numbers were associated with PMIDs via *PubMed links* only, 7 via *PubMed Central links* only, and 36 via both. In workflow 3, 15 accession numbers were associated with PMIDs via *PubMed links* only, none via *PubMed Central links* only, and 30 via both.

Agreement between workflows. Agreement between workflows was assessed regarding the accession numbers for which they retrieved PMIDs. The agreement between workflows 1 and 2 ($K = 0.388$, $P < 0.001$) and between 2 and 3 ($K = 0.340$, $P < 0.001$) was slight. Workflows 2 and 3 showed good agreement ($K = 0.791$, $P < 0.001$).

Table 2
Assessment of agreement between workflows

	Comparison 1: nucleotide accession numbers associated with one or more PubMed IDs	Comparison 2: PubMed IDs retrieved	Comparison 3: nucleotide sequence accession number–PubMed ID pairs retrieved
1 and 2: Cohen's *kappa*	0.388*	0.500*	0.242*
Level of agreement	Slight	Fair	Slight
1 and 3: Cohen's *kappa*	0.791*	−0.159*	−0.060
Level of agreement	Good	No agreement	
2 and 3: Cohen's *kappa*	0.340*	−0.305*	−0.636*
Level of agreement	Slight	No agreement	No agreement

* Indicates P value < 0.001. Cohen's kappa statistics calculated using SPSS 11.5.0. [19]; significance levels reported after application of the Bonferroni correction for multiple significance tests [20]. Interpretation of kappa statistic, per Byrt [18]: ≤ 0 = No agreement; 0.01 to 0.20 = Poor agreement; 0.21 to 0.40 = Slight agreement; 0.41 to 0.60 = Fair agreement; 0.61 to 0.80 = Good agreement; 0.81 to 0.92 = Very good agreement; 0.93 to 1.00 = Excellent agreement.

Comparison 2: Which and how many PMIDs were retrieved using the different workflows?

Overall results. A total of 338 PMIDs were retrieved: 72 by workflow 1, 101 by workflow 2, and 267 by workflow 3. Thirty-nine PMIDs were retrieved by all 3 workflows.

***PubMed links* and *PubMed Central links* paths.** Workflow 2 retrieved 56 PMIDs via *PubMed links* only, 36 via *PubMed Central links* only, and 9 via both. In workflow 3, 250 PMIDs were retrieved via *PubMed links* only, none via *PubMed Central links* only, and 17 via both.

Agreement between the workflows. Agreement between workflows was assessed regarding which PMIDs they retrieved. The agreement between workflows 1 and 2 was fair ($K = 0.500$, $P < 0.001$). There was no agreement between workflows 1 and 3 ($K = -0.159$, $P < 0.001$) or between workflows 2 and 3 ($K = -0.305$, $P < 0.001$).

Comparison 3: Which and how many accession number–PMID pairs were produced using the different workflows?

A workflow results in an accession number–PMID pairing when inputting the accession number to the workflow retrieves the PMID. The purpose of the workflows here was to retrieve literature on the function of the genes associated with each of the accession numbers; therefore, the accession number–PMID pairings were of particular interest.

Overall results. A total of 464 distinct accession number–PMID pairs were retrieved: 73 from workflow 1, 192 from workflow 2, and 301 from workflow 3. Overlap between the 3 workflows was fairly low, including 39 pairs resulting from all 3 workflows.

***PubMed links* and *PubMed Central links* paths.** In workflow 2, 117 accession number–PMID pairs resulted from the *PubMed links* only, 65 resulted from the *PubMed Central links* only, and 10 pairs resulted from both paths. In workflow 3, 254 pairs resulted from the *PubMed links* only, none resulted from the *PubMed Central links* only, and 47 resulted from both paths.

Agreement between the workflows. Agreement between workflows was assessed regarding which accession number–PMID pairings they produced. The agreement between workflows 1 and 2 was slight ($K = 0.242$, $P < 0.001$). There was no agreement between workflows 2 and 3 ($K = -0.636$, $P < 0.001$), and the comparison between workflows 1 and 3 was not statistically significant.

DISCUSSION

The results show the three workflows are neither strictly equivalent nor even nearly equivalent in the sense of strong agreement or overlapping of results. The significant differences among the workflows

Table 3
Comparison of alternate paths within workflows

	Comparison 1: nucleotide accession numbers associated with one or more PubMed IDs	Comparison 2: PubMed IDs retrieved	Comparison 3: nucleotide sequence accession number–PubMed ID pairs retrieved
In workflow 2, number retrieved by:			
PubMed links path only	83	56	117
PubMed Central links path only	7	36	65
Both paths	36	9	10
In workflow 3, number retrieved by:			
PubMed links path only	15	250	254
PubMed Central links path only	0	0	0
Both paths	30	17	47

might surprise an otherwise scientifically sophisticated user who is not an expert in the use of these information resources.

In this case, the existing Help documentation for the information resources can account for differences in the workflow output. The PubMed Secondary Source or SI field documentation accounts for differences between workflows 1 and 2. According to PubMed's Help information [13], the SI field and the PubMed links to GenBank are generated differently and are themselves not linked. The SI field identifies GenBank accession numbers discussed in MEDLINE articles, while the GenBank reference field (which for a given record includes citations that discuss the associated sequence) is used to create the PubMed links to GenBank. The Entrez Gene documentation accounts for differences between workflow 3 and workflows 1 and 2. The Entrez Gene PubMed Links documentation indicates that some Entrez Gene PubMed links are generated from GeneRIFs, as indicated by the PubMed (GeneRIF) option [21], and that the GeneRIF mechanism is a way to let scientists themselves add to the functional annotation of genes [22].

Although such documentation is available, the biologist using or designing workflows may not know about it. It is no more reasonable to expect biologists to be experts in the metadata of biological information resources than it is to expect librarians to be experts in biology. Thus, because even simple, apparently similar information retrieval workflows may produce different results, a multidisciplinary team approach to authoring, vetting, and using life science workflows is needed. Such teams must include experts in the primary science and experts in the metadata characterizing the information resources.

The importance of librarians as metadata experts in life science research was recognized by the Human Genome Project in 1997 [23]. Unfortunately, almost a decade later, the library remains largely excluded from the mainstream of life science research: very few universities offer bioinformatics end-user support services through the library [24]; demand is generally not great for such services when offered [25]; and molecular biology students in particular do not choose the library as their preferred source of information about bioinformatics databases [26].

The life science information space is growing extremely rapidly, largely facilitated by "the breakdown of the traditional barriers between academic disciplines and the application of technologies across these disciplines" [27]. Similarly, breaking down the barriers between "scientist" and "librarian" and fostering the interdisciplinary and synergistic combination of their respective expertise in the development and use of life science workflows are crucial to achieving full and optimal exploitation of the life science information space.

REFERENCES

1. EverythingBio. Definition of life science. [Web document]. EverythingBio.com. [cited 25 Jan 2007]. <http://www.everythingbio.com/glos/definition.php?word=life+science>.
2. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T. Taverna: a tool for building and running workflows of services. Nucleic Acids Res 2006 Jul 1;34(Web server issue):W729–W732.
3. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 2004 Nov 22;20(17):3045–54.
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2005 Jan 1;33(database issue):D34–D38.
5. US National Library of Medicine, National Center for Biotechnology Information. PubMed overview. [Web document]. Bethesda, MD: The Library. [rev. 30 Jun 2006; cited 14 Aug 2006]. <http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html>.
6. Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. Nucleic Acids Res 2006 Jul 1;34(Web server issue):W6–W9.
7. Masys DR. Linking microarray data to the literature. Nat Genet 2001 May;28(1):9–10.
8. US National Library of Medicine, National Center for Biotechnology Information. Databases. [Web document]. Bethesda, MD: The Library. [rev. 17 Jan 2006; cited 1 Dec 2006]. <http://www.ncbi.nlm.nih.gov/Database/>.
9. Xuan W, Watson SJ, Meng F. GeneInfoMiner—a Web server for exploring biomedical literature using batch sequence ID. Bioinformatics 2005 21(16):3452–3.
10. Pattison JS, Folk LC, Madsen RW, Childs TE, Booth FW. Transcriptional profiling identifies extensive downregulation of extracellular matrix gene expression in sarcopenic rat soleus muscle. Physiol Genomics 2003 Sep 29;15(1):34–43.
11. Pattison JS, Folk LC, Madsen RW, Childs TE, Spangenburg EE, Booth FW. Expression profiling identifies dysregulation of myosin heavy chains IIb and IIx during limb immobilization in the soleus muscles of old rats. J Physiol 2003 Dec 1;553(pt 2):357–68.
12. Pattison JS, Folk LC, Madsen RW, Booth FW. Selected contribution: identification of differentially expressed genes between young and old rat soleus muscle during recovery from immobilization-induced atrophy. J Appl Physiol 2003 Nov;95(5):2171–9.
13. US National Library of Medicine, National Center for Biotechnology Information. PubMed help. [Web document]. Bethesda, MD: The Library. [rev. 8 Aug 2006; cited 14 Aug 2006]. <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.chapter.pubmedhelp>.
14. US National Library of Medicine, National Center for Biotechnology Information. Entrez nucleotide. [Web document]. Bethesda, MD: The Library. [rev. 17 Jan 2006; cited 14 Aug 2006]. <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide>.
15. US National Library of Medicine, National Center for Biotechnology Information. PubMed Central overview. [Web document]. Bethesda, MD: The Library. [rev. Jan 7 2005; cited Aug 17 2006]. <http://www.pubmedcentral.nih.gov/about/intro.html>.
16. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2005 Jan 1;33(database issue):D54–D58.
17. Patrick T, Folk LC, Craven CK. Asymmetries in retrieval of gene function information. Presented at: MLA '06, 106th Annual Meeting of the Medical Library Association, "Transformations A-Z"; Phoenix, AZ; May 19–24, 2006.
18. Byrt T. How good is that agreement? Epidemiology 1996 Sep;7(5):561.
19. SPSS. SPSS for Windows, release 11.5.0. SPSS, 2002.

20. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995 Jan;310(6973):170.
21. US National Library of Medicine, National Center for Biotechnology Information. Entrez Gene help: integrated access to genes of genomes in the reference sequence collection: finding data related to Entrez gene in other Entrez databases. [Web document]. Bethesda, MD: The Library. [rev. 13 Nov 2006; cited 30 Nov 2006]. <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpgene.section.EntrezGene.Finding_Data_Related>.
22. US National Library of Medicine, National Center for Biotechnology Information. GeneRIF—Gene Reference Into Function. [Web document]. Bethesda, MD: The Library. [rev. 30 Nov 2006; cited 30 Nov 2006]. <http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html>.
23. Human genome project [report no.] JSR-97–315. [Web document]. The Mitre Corporation. [rev. 1997; cited 29 Jan 2007]. <http://www.ornl.gov/sci/techresources/Human_Genome/publicat/miscpubs/Jason/>.
24. Messersmith DJ, Benson DA, Geer RC. A Web-based assessment of bioinformatics end-user support services at US universities. J Med Libr Assoc 2006 Jul;94(3):299–305,E156–E187.
25. Geer RC. Broad issues to consider for library involvement in bioinformatics. J Med Libr Assoc 2006 Jul;94(3):286–98,E152–E155.
26. Brown C. Where do molecular biology graduate students find information? Sci Technol Libr 2005; 25(3):89–104.
27. Welsh E, Jirotka M, Gavaghan D. Post-genomic science: cross-disciplinary and large-scale collaborative research and its organizational and technological challenges for the scientific research process. Philos Trans Roy Soc Lond A 2006 Jun 15;364(1843):1533–49.

AUTHORS' AFFILIATIONS

Timothy B. Patrick, PhD (corresponding author), tp5@uwm.edu, Assistant Professor, College of Health Sciences, University of Wisconsin–Milwaukee, Milwaukee, WI; **Catherine K. Craven, MLS, MA,** catherine.craven@gmail.com, Department of Health Management and Informatics, School of Medicine; **Lillian C. Folk, MS,** folkl@missouri.edu, College of Veterinary Medicine, University of Missouri–Columbia

Received September 2006; accepted December 2006

Appendix B

Members of the Life Sciences Research Task Force:

Kate Anderson (Chair), Specialized Services Librarian, J. Otto Lottes Health Sciences Library; Zalk Veterinary Medical Library

C. Trenton Boyd, Head, Zalk Veterinary Medical Library

Catherine Craven, Bioinformatics Librarian, Welch Library, Johns Hopkins University

Janice Dysart, Science Librarian, Ellis Library

Brenda Graves-Blevin, Science Librarian, Ellis Library

E. Diane Johnson, Head of Information Services, J. Otto Lottes Health Sciences Library

Mary Ryan, Head of Reference, Ellis Library

Caryn Scoville, Head of Interlibrary Loan, J. Otto Lottes Health Sciences Library

MaryEllen Cullinan Sievert, Consultant, J. Otto Lottes Health Sciences Library

Chris Topinka, Doctoral Student in Biomedical Informatics (Computer Science)

Deborah H. Ward, Director, J. Otto Lottes Health Sciences Library

Appendix C

Appendix D

September 13, 2007

To: Jim Cogswell, Director of MU Libraries

From: MU Libraries Life Sciences Research Task Force: Kate Anderson (Chair), Trenton Boyd, Janice Dysart, Brenda Graves-Blevins, Diane Johnson, Mary Ryan, Caryn Scoville, MaryEllen Sievert, Deb Ward

RE: **Bioinformatics and MU Libraries**

On July 30 and 31, 2007, Diane Rein, Assistant Professor of Library Science and Assistant Life Sciences Librarian at Purdue University, visited MU Libraries to conduct a six-hour introductory workshop on bioinformatics and to participate in a discussion on how MU Libraries can deliver bioinformatics services to life sciences researchers on campus. Dr. Rein's visit was supported by funding from the Institute of Museum and Library Services through a "Bring in the Expert Grant" administered by the Missouri State Library.

The "Introducing Bioinformatics: A Primer for Librarians" workshop was attended by fourteen librarians from MU, UMR, UMKC and Washington University as well as a biological sciences faculty member and the life sciences consultant currently working with MU Libraries. Based on a survey of participants, the workshop received high marks for the presentation, presenter, and accompanying manual.

Fifteen people participated in the discussion session the following morning. Besides Diane Rein, the attendees included librarians and administrators from MU Libraries as well as the life sciences consultant for MU Libraries, a representative from the MU Division of Information Technology, a faculty member from the Department of Computer Science and the Director of the Bond Life Sciences Center. Participants concluded that currently there is a fragmented approach to meeting bioinformatics needs on campus with several "information silos" having little interaction.

From the consultation session and further discussion, a number of possibilities for the MU Libraries to develop new bioinformatics services emerged:

- Serve as a facilitator or clearinghouse to bring the different bioinformatics factions/information silos together.
- Hire a full-time Bioinformatics Librarian. While the Life Sciences Research Task Force realizes the current financial situation on campus may prevent the creation of a new position, we will draft a position description that reflects local need.
- Pursue partnerships with Chi-Ren Shyu, Associate Professor of Computer Science, and others in a number of areas: e.g., develop internships or fellowships in bioinformatics for graduate students in computer science; submit grants on bioinformatics projects of interest; etc.

As a next step, Jack Schultz, Director of the Bond Life Sciences Center, proposed that MU Libraries produce a white paper explaining the new role MU Libraries could take in this area and how we could interact with other units on campus to enhance bioinformatics services. The Life Sciences Research Task Force expects that the process of developing this white paper will help us delineate our capabilities and possible contributions to bioinformatics on the MU campus.

Appendix E

DRAFT (10/04/07)

Molecular Biology/Bioinformatics Specialist
University of Missouri-Columbia Libraries (MU Libraries)

Position Description: Reports to the Director of the Health Sciences Libraries, with dotted line reporting to the Director of the Bond Life Sciences Center. Molecular Biology/Bioinformatics Specialist develops and conducts a program for supporting biosciences and bioinformatics programs across the University of Missouri-Columbia campus.

Responsibilities:

- Develop and teach tailored education sessions related to the effective use of specialized bioinformatics and molecular biology databases and information resources.
- Create and maintain exemplary molecular biology and bioresearch information resources, including web-based information access tools and information portals.
- Consult and collaborate with researchers to address specific technical and research issues, based on user needs assessment and evaluation.
- Partner with campus librarians regarding educational and reference services
- Participate in scholarly and service activities of the MU Libraries and the University of Missouri-Columbia

Qualifications:

- Advanced degree in molecular biology, genetics, or related science. Master's degree in Library Science from an ALA-accredited school.
- Six or more years' related professional experience. Proven knowledge of principles, theories, practices, terminology, and research trends in genetics, molecular biology, bioinformatics and related disciplines.
- Skilled in the use and manipulation of molecular resources, software and search engines, including sequence, structure, proteomic and genomic resources.
- Experience in developing and delivering training on a variety of molecular biology resources
- Experience creating web-based information tools. Demonstrated willingness to embrace new and emerging technologies
- Team-oriented, flexible, and able to work both independently and collaboratively in a complex, rapidly changing environment.
- Evidence of continued professional growth
- Excellent oral and written communication skills

Selected Bibliography

Brown C**.** The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 54 (10): 926-938 AUG 2003.

Brown C. Where do molecular biology graduate students find information? SCIENCE & TECHNOLOGY LIBRARIES 25 (3): 89-104 2005.

Geer RC. Broad issues to consider for library involvement in bioinformatics. JOURNAL OF THE MEDICAL LIBRARY ASSOCIATION 94 (3): 286-298 JUL 2006.

Geer RC, Rein DC. Building the role of medical libraries in bioinformatics. JOURNAL OF THE MEDICAL LIBRARY ASSOCIATION 94 (3): 284-285 JUL 2006.

MacMullen WJ, Denn SO. Information problems in molecular biology and bioinformatics JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 56 (5): 447-456 MAR 2005.

Messersmith DJ, Benson DA, Geer RC. A Web-based assessment of bioinformatics end-user support services at US universities . JOURNAL OF THE MEDICAL LIBRARY ASSOCIATION 94 (3): 299-305 JUL 2006.

Patrick TB, Craven CK, Folk LC. The need for a multidisciplinary team approach to life science workflows. JOURNAL OF THE MEDICAL LIBRARY ASSOCIATION 95 (3): 274-278 JUL 2007.