Medical Library Association Comments to the National Library of Medicine in response to NLM's Request for Information on Next Generation Data Science Challenges in Health and Biomedicine

Submitted October 31, 2017

Background/Information Requested: *To continue to strengthen and expand the scope of biomedical data science research, NLM seeks new recommendations from public and private organizations, industry and individuals about priorities for the next phase of NIH investment in biomedical data science. In particular, recommendations are invited in any or all of the focal areas outlined below, but additional comments reaching beyond these areas are also welcomed.*

MLA's comments addressed Section 3 of the RFI.

SECTION 3:
Promising directions for workforce development and new partnerships:

5. Workforce Development and Diversity:

In the era of data-powered health, it is essential that research data is findable, accessible, interoperable, and reusable (FAIR), (https://www.force11.org/group/fairgroup/fairprinciples) and maximized to benefit health care. Researchers also must follow the principles of reproducible research (http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124). Health sciences librarians are well-suited and well-positioned to work with researchers to achieve these goals. Dr. Patricia Brennan, National Library of Medicine (NLM) Director, emphasized the need for increased capacity for data support services among health sciences libraries in her 2017 MLA/NLM Leiter Lecture. Specialized skills in data management are now essential for health sciences information professionals, and the competencies for several organizations, including the Medical Library Association's Competencies for Lifelong Learning and Success (MLA) (http://www.mlanet.org/p/cm/ld/fid=1217), have been updated to reflect the importance of this growing area of expertise. Specifically, MLA's competency #5 (evidence-based practice and research) expects that "a health information professional curates and makes accessible bioscience, clinical, and health information data, information and knowledge."

Throughout its recent strategic planning process (https://www.nlm.nih.gov/pubs/plan/strategic_planning.html), NLM has recognized its role to advance data science, and the need to continue providing education and training opportunities for its own staff, for biomedical librarians through the National Network of Libraries of Medicine (NNLM), and in cooperation with professional associations, for health professions students, and for practicing clinicians and researchers. The needs of all of these groups will be considered in the formulation of NLM's recommendations related to workforce development.

Several organizations, including MLA, currently are gathering information that will help identify areas of knowledge and gaps within the knowledge base of the health sciences library community. We recommend that MLA and NLM collaborate to evaluate this information and to develop curriculum and professional development opportunities. We also recommend providing partnership opportunities for the health sciences library community and the NNLM Evaluation Office (NEO) to evaluate the impact of these courses and training programs.

While there are several self-directed learning opportunities for librarians in the area of data science, there is much room for expanded educational formats, focused community building, and structured training designed specifically for information professionals. The Oregon Health & Science University (OHSU) offers open educational BD2K training modules (https://dmice.ohsu.edu/bd2k/) and more self-directed learning can be discovered through the BD2K Training Center's educational resource discovery index (ERuDIte) (https://bigdatau.ini.usc.edu/about_erudite). MLA offers a few continuing education (CE) webinars around data visualization (http://www.medlib-ed.org/all-courses) and a symposium on reproducibility (http://www.mlanet.org/p/cm/ld/fid=1173), but the association is continually building up its selection of data-centric learning opportunities taught by data librarians within the membership.

As an example of a robust training program, in August 2017 NLM awarded the NNLM Training Office (NTO) an administrative supplement to develop Biomedical and Health Research Data Management Training for Librarians (https://news.nnlm.gov/nto/2017/09/20/a-new-training-program-biomedical-and-health-research-data-management-for-librarians/). While there are many resources available to learn about data management principles and services, there is a need for a comprehensive training program that brings together the best of these resources and enhances them with meaningful, practical activities focused on biomedical and health research data. This new training program builds on existing resources and transforms the learning experience from a largely self-directed, isolated endeavor to an organized program that is supported by experienced peer mentors, many of whom belong to the MLA Data Special Interest Group. Course topics include an overview of data management, choosing appropriate metadata descriptors or taxonomies for a dataset, addressing privacy and security issues with data, and creating data management plans. Each module will be co-taught by a practicing data librarian, and the experience will culminate in a Capstone Summit at NIH, where participants can demonstrate improved skills and knowledge

through a capstone project and help evaluate the program for future cohorts. MLA CE credit will be awarded to those who complete the entire program. Participants in this program will not only be prepared to support research data management at their institutions, but also ready to take next steps to support data science and reproducible research.

While initiatives like this are a great start, more structured education and training programs are needed within the health sciences library community. The NLM Biomedical Informatics Course offered an outstanding training opportunity for health sciences librarians and information professionals. MLA would be interested in partnering with NLM to develop the next iteration of this course as it relates to data management and data science. MLA also recommends partnering to support health sciences librarians, who are spearheading educational efforts around reproducible research (ex: http://www.mlanet.org/page/ce-symposium (http://www.mlanet.org/page/ce-symposium) and https://github.com/shirl0207/reproducible_science) (https://github.com/shirl0207/reproducible_science). In addition, we propose to build on all these efforts with the recommendation for the establishment of an MLA/NLM Data Science Specialization, which is described in more detail in the next section of our comments.

Providing these new and emerging training programs will strengthen and build diversity of professional competencies and roles for health sciences librarians, which aligns with MLA's newly established diversity goal to evaluate and improve MLA practices as they relate to diversity and inclusion within the association. It also supports NLM's diversity goal to have a strong and diverse workforce in the areas of research and professional practice. With this shared commitment to diversity, MLA hopes to work more closely with NLM to promote strategic partnerships (further addressed in the next section).

6. New Stakeholder Partnerships:

I. The Medical Library Association (MLA) proposes developing a Data Science Specialization in partnership with NLM to meet the continuing education needs of its membership, as well as augment the current set (http://www.mlanet.org/p/cm/ld/fid=42). What follows is a justification for why the new specialization will be of value to health sciences librarians and other information professionals.

In 2017, MLA administered a questionnaire about skills necessary for each of its professional competencies (http://www.mlanet.org/p/cm/ld/fid=1217). Over 400 respondents rated their skills as none, basic, or expert. Three items are of interest for this RFI because the responses broadly indicate gaps in data science skills:

1. *Conserving, preserving, and archiving print and digital materials.* Half of respondents reported no skills. This is of concern and indicates gaps in traditional MLS educational programs. In today's digital ecosystem, this skill set is closely related to retrieval and

management of institutional assets. Thus, librarians need skills in knowledge representation, metadata annotation, semantic analysis, and assessment. The MLA Data SIG also identified metadata as an important topic for future training. Metadata is critically important for data curation, preservation, and sharing.

The Center for Expanded Data Annotation and Retrieval (CEDAR). A potential strategic partner in developing a module for a Data Science Specialization could be CEDAR, which is one of the Big Data to Knowledge centers (https://metadatacenter.org). Its mission is to optimize "the metadata pathway from provider to end user." It would seem a natural fit for MLA to partner with CEDAR.

2. *Implementing data management plans (DMPs).* About half of respondents reported no skills. The majority of respondents could not conduct a data curation interview, develop and implement a DMP, or consult on managing data across the data life cycle. A potential training module to address this gap is described under Workforce Development. The MLA Data SIG also identified the following as important topics for training: (1) evaluation of DMPs and (2) how to talk to scientists about their data. Communication with scientists and researchers will increasingly require proficiency in this domain.

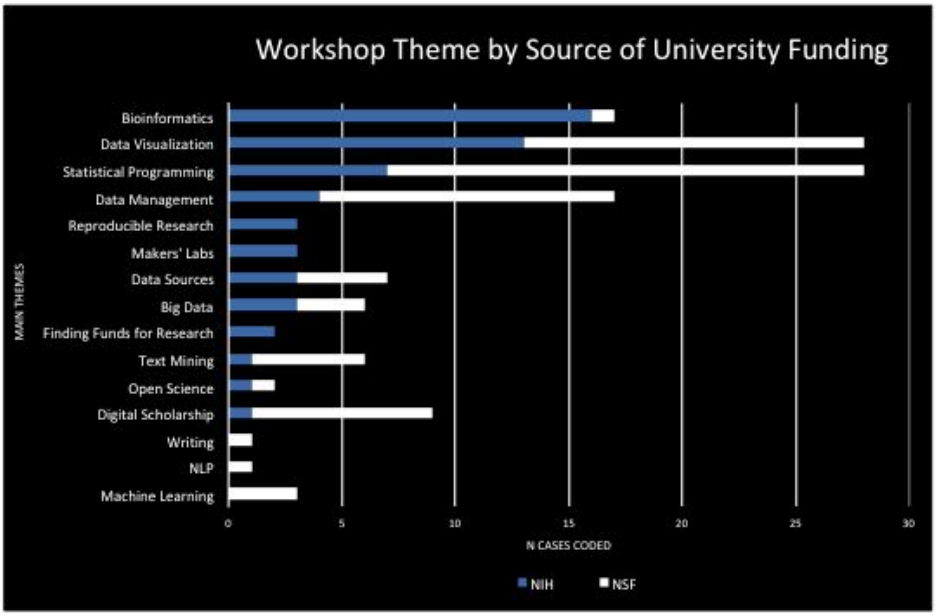
3. *Interpreting data and presenting statistical and data analyses.* About a third of respondents reported no skills. Addressing this gap would mean developing modules on: (1) analysis of mixed data (heterogeneous data types, such as audio, visual, geographic, numeric, and textual); (2) interpretation of results, paying attention to assumptions underlying the methods of analysis; (3) display of findings for exploration and presentation; and (4) writing up results given known standards for particular study designs and editorial expectations, e.g. those of the International Committee of Medical Journal Editors (http://www.icmje.org/about-icmje). The MLA Data SIG identified the following as important topics related to this competency:
    a. Visualizing data, especially using R and Tableau
    b. Statistical programming in R and Python
    c. Developing relational databases
    d. Training in REDCap, a Web application for building and managing online surveys and databases
    e. Working with open source tools
    f. Measuring research impact

    Potential partners could involve developers of MOOCs and Open Educational Resources. More than likely, an environmental scan of resources that could be revised for librarians should be conducted.

II. MLA suggests developing strategic partnerships with federal agencies in addition to NIH, such as NSF and the National Endowment for the Humanities (NEH). The latter could help develop training modules in the digital humanities relevant to social and behavioral medicine. Additionally, multi-disciplinary partnerships would be in keeping with team science and better

reflects the federal funding environment. For example, NSF partners with a variety of federal agencies and organization (https://www.nsf.gov/about/partners/fedagencies.jsp), including NIH.

III. In 2017, Bekhuis analyzed workshops in support of data-driven research offered by libraries in top NIH- and NSF-funded universities [1]. Coverage of data visualization (including geographic information systems), data sources and big data was about the same across funding subsets, whereas coverage of writing, natural language processing (NLP), and machine learning was unique to the NSF-funded subset. In a finer-grained analysis, statistical topics included various languages and software, such as R, Python, and SPSS, and various platforms and tools, such as the Natural Language Toolkit (NLTK) and MAXQDA for qualitative data analysis, among others. Closely related to machine learning topics, were text mining, NLP, and corpus linguistics. The identified themes could guide development of modules for the Data Science Specialization. Additionally, by covering topics offered solely by libraries in NSF-funded universities, MLA could help their members improve support of multi-disciplinary research.



Source. Bekhuis (2017). Library Workshops in Support of Data-Driven Research in Top NIH and NSF-Funded Universities.

Conclusion

The MLA suggestions for developing a new Data Science Specialization in partnership with NLM, as well as suggestions for other strategic partnerships, are based on evidence derived from MLA members and the health sciences literature. However, the ideas presented here also correspond with evidence from other groups, such as the most recent environmental scan by

the Association of College and Research Libraries (ACRL) [2], the ACRL report of top trends in academic libraries [3], the federal strategic plan for big data research and development [4], a content analysis of strategic plans in academic research libraries [5], and an effort funded by the Institute of Museum and Library Services (IMLS) to guide implementation of data science in libraries [6]. Overall, the cumulative evidence from various domains appears to corroborate the need for a Data Science Specialization.

References

1. Bekhuis T, EDDA Analytics Group™. Library Workshops in Support of Data-driven Research in Top NIH- and NSF-funded Universities. TCB Research & Indexing LLC. No. 00170205v2. February 2017. http://www.tanjabekhuis.com/wp-content/uploads/2017/08/Bekhuis_Library-Workshops-NIH-and-NSF_REPORT-v2.pdf
2. Association of College and Research Libraries (ACRL) Research Planning and Review Committee. Environmental Scan 2017. March 2017. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/EnvironmentalScan2017.pdf
3. Association of College and Research Libraries (ACRL) Research Planning and Review Committee. 2016 Top Trends in Academic Libraries: A Review of the Trends and Issues Affecting Academic Libraries in Higher Education. C&RL News. June 2016. http://crln.acrl.org/index.php/crlnews/article/view/9505/10798
4. Big Data Senior Steering Group, Networking and Information Technology Research and Development Program, National Science and Technology Council, Executive Office of the President. The Federal Big Data Research and Development Strategic Plan. May 2016.
5. Saunders L. Academic Libraries' Strategic Plans: Top Trends and Under-Recognized Areas. The Journal of Academic Librarianship. 2015;41(3):285-91.
6. Burton M, Lyon L. Data Science in Libraries. Bulletin of the Association for Information Science and Technology. 2017;43(4):33-35.