MLA Medical Library Association

AAHSL Association of Academic Health Sciences Libraries

Comments of the Association of Academic Health Sciences Libraries and Medical Library Association Re: RFI: Input into the Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics

Submitted March 12, 2012 via http://grants.nih.gov/grants/guide/rfi_files/nih_di/add.cfm

These comments are submitted on behalf of the Medical Library Association (MLA) and Association of Academic Health Sciences Libraries (AAHSL) and address the following issues: standards development, secondary/future use of data, data accessibility, incentives for data sharing, and support needs.

The Association of Academic Health Sciences Libraries (AAHSL) (http://www.aahsl.org) is composed of the directors of the libraries of 116 accredited U.S. and Canadian schools as well as 28 associate members. AAHSL's goals are to promote excellence in academic health sciences libraries and to ensure that the next generation of health practitioners is trained in information seeking skills that enhance the quality of healthcare delivery.

The Medical Library Association (MLA) (http://www.mlanet.org) is a nonprofit educational organization with approximately 4,000 health sciences information professional members worldwide. Founded in 1898, MLA provides lifelong educational opportunities, supports a knowledgebase of health information research, and works with a global network of partners to promote the importance of quality information for improved health to the health care community and the public.

Standards Development

AAHSL and MLA believe that work is definitely needed in developing and sharing data standards, definitions, and ontologies. Researchers are struggling to create their own approaches or trying to use definitions and structures intended for fields outside their own, for example, using CaBIG resources for surgical research.

NIH, through the National Library Medicine (NLM), has played an important role in identifying standards for clinical care and knowledge. It would be beneficial to have that expertise applied to the research world. This would require additional funding support for NLM to take on this role, but it is one where NLM has the experience and expertise that would benefit the entire research community.

Additional standards must be developed for those datasets that are not digital. Biological specimens, MRI images, ultrasound videos and other formats need the same approach in terms of commonly used data definitions, standards for identifying methodologies, and controlled vocabularies of terms.

Librarians can be essential team players, not only in helping to develop standards and ontologies, but also in making their research communities aware of the resources available through NIH and other research groups and agencies.

Secondary/future use of data

In order to share data in the future, researchers must have commonly defined data fields with specified structures for that data, and standard definitions for methodologies that can be linked to that data. These approaches will ensure not only that it can be shared, but that it will be meaningful and relevant for use in the future and by others working on the same project.

Data standards and definitions will also allow other experts to review and evaluate the data to ensure that it is valid and replicable. Other disciplines less familiar with a specific research area would be able to repurpose these datasets, leveraging the data collected and the funding used to support the original research. Once these common operational standards are in place, it should be possible to more easily extract data and present it to the general public, supporting research findings and outcomes.

These standardized approaches to large datasets would also enable future IRB (Institutional Review Board) review of older datasets when researchers want to repurpose the data or conduct retrospective studies. When the content of datasets or fields is not known, the risk for releasing or inappropriately using Personal Health Identifiers (PHI) exists, and it will be difficult for IRBs to judge the risks using shared data for research.

Since research is diverse it is unlikely that standards will address every possible data collection need. Therefore, guidelines on best practices for creating and documenting data points should also be developed so that biomedical libraries and research teams can work together on the specific needs of a particular project or lab, without jeopardizing the long-term usefulness of the dataset to the researchers or others.

At the very least, it would be useful to create a repository of data structures, definitions, ontologies, etc. that have been developed by government agencies, organizations and research institutions so that researchers could begin to use structures that have proven useful in a research setting and avoid recreating the wheel. This could be an “open source” collection of “standards” generated, revised, and used by the research community.

Again, AAHSL and MLA maintain that librarians have the skills and expertise to assist researchers in understanding the necessity for, and applying the criteria for data definitions so that it can be shared in the future. Librarians can play an important role from the early planning of research proposals to the implementation of data management once a project is funded and should be part of the research team.

Another related issue is accessing paper records with older datasets containing PHI. More guidance is needed as to when older data sets can be accessed and shared, for example what are the requirements or review standards for research using records containing patient data sets prior to 1950, 1900, etc. Some historical medical records are of interest to researchers who study populations, trends in diseases and conditions, and public health issues. Many of these records are currently unavailable because institutions are concerned by the presence of PHI and how IRBs can review paper records that may not be standardized in terms of data or its presentation. Additional guidelines on older data sets would make these resources more easily accessible to investigators.

Data Accessibility

A central repository of research data would increase sharing and leverage other research that builds off existing datasets. If feasible, this should be a long-term goal for NIH. However, this is a tremendous undertaking and many datasets that are not federally funded may be excluded from such an approach. Another approach is to create a central indexing repository where information and links to other data repositories resides. Basic information about data definitions, methodologies, and ontologies, in abstract form, could be submitted by researchers along with other key information. Such a central index or clearinghouse would enable researchers to locate other datasets and make their own work more visible and accessible.

Incentives for data sharing

Standards for reporting data citations and data publications are needed so that attribution is given to the original creators of data and to enable the tracking of the impact or usefulness of the data to other research endeavors. This would enable the inclusion of shared data activities as part of annual faculty evaluation and tenure and promotion review, similar to the current practice of citations for peer-reviewed journal articles. If the use of datasets is clearly acknowledged and cited, researchers could include as part of their faculty portfolio documentation on the extent to which their research data have been actually used (cited) or potentially used (published) in producing other research.

Data peer-review mechanisms could also create an incentive for producing high quality and shareable datasets. There would also need to be mechanisms to distinguish and differentially weight peer-reviewed data citations in relation to datasets that are simply made publicly available, but have less impact or usability within the research community.

Support Needs

AAHSL and MLA believe that more training needs to be provided on data curation and management, but at several levels. Certainly complex research needs more individuals trained in computational and analytical methods and there should be more funding for fellowships.

Additionally, there are other staff members within institutions who would benefit in training programs for the curation and management of data.

Librarians already are working with students and faculty on related research issues, and further training would enable them to train the research community in best practices, along with helping them to understand the importance of managing and sharing their datasets. In partnership with computational bio-informaticists and statisticians, librarians undertaking additional training opportunities can address data stewardship principles and practices including: data archival methods; metadata creation and usage; and awareness of storage, statistical analysis, archives and other available resources as part of a data stewardship training curriculum. Librarians, in partnership with other disciplines and experts can support training of the future research investigators and the workforce. It is recommended that the NIH develop training programs and fellowships that further develop the skills of this workforce that already exists in most institutions.