Provost

Northwestern University
633 Clark Street
Evanston, Illinois 60208-1101

Phone 847-491-5117
Fax 847-467-1630
nu-provost@northwestern.edu

NORTHWESTERN
UNIVERSITY

January 12, 2012

The Honorable John P. Holdren
Assistant to the President for Science and Technology and
Director, Office of Science and Technology
New Executive Office Building
725 – 17th Street, NW
Washington, DC 20502

Comments in response to Office of Science and Technology Policy Request for Information: Public Access to Digital Data Resulting From Federally Funded Research
Federal Register Doc No 2011-28621
http://www.gpo.gov/fdsys/pkg/FR-2011-11-04/html/2011-28621.htm

Dear Dr. Holdren:

We are grateful to the Office of Science and Technology Policy for the opportunity to submit comments about providing public access to research data. Rather than a point-by-point response to the questions in the Request for Information, we offer these general comments about the challenges and benefits of data sharing, and issues the federal government should consider in seeking to implement new requirements.

Northwestern University is a private research institution with 16,377 students and approximately 3,000 full time faculty. In academic year 2010-11, Northwestern researchers attracted total awards and grants of approximately $511.7 million. Northwestern's libraries hold more than 5 million volumes, 4.6 million microforms, and provide access to 110,341 current periodicals and serials. In addition, the library system boasts more than 700 databases and 6,000 electronic journals. 56% of the libraries' $14 million collections budget is devoted to these e-resources.

Northwestern is recognized both nationally and internationally for the quality of its educational programs at all levels. *U.S. News & World Report* consistently ranks the University's undergraduate programs among the best in the country.

Among graduate programs, the Kellogg School of Management regularly ranks among the top five business schools in the country for both its traditional curriculum and its executive master's program. *U.S. News & World Report* rankings placed Northwestern's School of Law 11th, and the Feinberg School of Medicine in the top 20.

Sharing and Public Accessibility

The absence of a policy requiring investigators to take specific action to share and preserve research data has resulted in management practices that are idiosyncratic and incomplete, which all but guarantees future data loss. Many Northwestern researchers already share their data with colleagues,

but exchanges may be informal or involve temporary sharing mechanisms (email attachments, temporary FTP servers, etc.) that do not take reuse or long-term preservation into consideration. While specific approaches will and should vary across disciplines, a policy that clearly articulates the definition and goals of providing public access to research data will support gradual development of standards and repository systems to enable responsible stewardship.

Current publication and preservation methods also often fail to identify clearly and consistently the data, data creators, and other provenance necessary to provide attribution in future work, or to address problems such as patent disputes. Researchers have an inadequate understanding of effective data management and curation practices. Many labs do not have the means to store permanently and provide internet access broadly to very large datasets, which may be petabytes in size. Therefore, a policy must be sensitive both to the significant technical challenges and the financial impact of sharing requirements. Centralized repositories for research data may prove to be a cost effective alternative that may alleviate investigators from the financial and technical burden of providing secure, reliable access to published results.

Perhaps most importantly, a policy should make clear what a public access requirement is designed to support. Reproducibility of research, independent verification of findings, and more rapid adoption of previous research to new investigations are all good reasons to mandate data sharing. Some data may not be sharable, either temporarily or permanently, for reasons of national security, privacy, or pending legal action, but these restricted data will still benefit from preservation services and application of standards to describe adequately and store safely research data. A clear statement of intent will help researchers determine whether all raw data, only significant findings, or only data directly linked to a publication are affected by a policy. It may be that an expansion of practice, if not policy, to encourage investigators to share negative results and other types of data not usually shared will also advance discovery.

A coordinated data sharing program must also clarify investigators' obligations to keep data safe, clearly define minimum acceptable practice for effective data management and curation, and tie compliance to ongoing funding. Careful consideration should be given to the design and development of tools that simplify metadata creation. Researchers who may be willing to share data will be very resistant to using awkward, poorly designed tools that disrupt active research or being forced to re-enter the same information repeatedly.

Copyright and Ownership

The legal status of research data must also be carefully considered. Facts do not satisfy the threshold for originality, and are therefore not eligible for protection under United States copyright law. 'Research data' is a broad term encompassing many different types of content, from the massive raw output of sensing instruments to text markup to painstakingly curated survey data and everything in between. While published research articles, survey instruments, software, and other research products will qualify for copyright protection, the data themself may not, so a different set of legal instruments may be needed to express the rights associated with data. Copyright transfer and licensing agreements, or the Creative Commons licenses that operate under a presumption of copyrightability, will not be sufficient to document the expectations of researchers. Open science and open data initiatives such as the Science Commons, specifically its database protocol project http://sciencecommons.org/resources/faq/database-protocol/, and the Panton Principles http://pantonprinciples.org/ provide a good discussion of data IP issues and examples of appropriate legal instruments. As with published research articles, a federal policy to promote public access to

data should not permit publishers to compel researchers to permanently restrict access to and use of their data as a condition of publication.

Funding and Implementation

Universities, their research administrators, libraries, and technology specialists are in a good position to advise investigators as they develop data management plans, and to help identify appropriate metadata standards, data description and normalization tools, and storage solutions. However, the costs of building these data management and preservation systems will be massive, and cannot be fully borne by individual universities.

The federal government has also struggled to maintain funding for large data storage projects, as demonstrated by the near closure of the Sequence Read Archive (SRA) in 2011. However, the SRA and other NCBI databanks, as well as those at the Food and Drug Administration and the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) are available to anyone worldwide and are well known throughout the research community.

Centralized or multi-organization approaches have the potential to be more cost-effective, create larger linked stakeholder communities both to advocate for continued funding and to develop standards, and increase success of data normalization efforts through provision of shared technology platforms. A hybrid approach of local infrastructure for active research phases, and centralized or multi-organization solutions for broad public sharing or long-term preservation is likely inevitable. These initiatives should emerge in parallel so that critical information about projects, software, algorithms, and other meta-information are consistently captured beginning early in the data lifecycle, and travel with the data to reduce barriers to sharing. If the government cannot provide centralized storage for research data, grant funding to researchers and their institutions must be increased to support expansion of local or disciplinary capacity, or to pay incremental costs associated with a single project.

Standards for versioning, selecting, describing and citing/attributing data must evolve in conjunction with the researcher communities who use them. Although far from comprehensive, here are a few comments and examples of current standards and development activities:

Citation, attribution and linking

If data are received directly from another researcher, new publications arising from these data must mention this in a methods section, and all data sets should be properly cited in methods and references sections (depending on norms for discipline and the specific journal's format). Failure to properly cite datasets should have the same consequences as other instances of plagiarism: retraction of manuscripts. The data must be cited consistently; see the DataCite project http://datacite.org/ for an example of a promising data set registry and identifier minting service. Implementing a data sharing and citation system is also a ripe opportunity for linked open data (LOD) and RDF to take the forefront. If each dataset has a unique identifier, it can be linked through RDF triple format (e.g. "Paper [paper identifier] has related dataset Y [dataset identifier]"), further enforcing consistency, but also significantly improving machine readability.

Standards for interoperability and reuse

The FDA is evaluating similar standards to MIAME (Minimum Information About Microarray Experiments) for ChIP-Seq and RNA-Seq data descriptions. The Gene Ontology project is an initiative with the aim of standardizing the representation of gene and gene product attributes across

species and databases. The project provides a controlled vocabulary of terms http://www.geneontology.org/GO.downloads.ontology.shtml for describing gene product characteristics and gene product annotation data http://www.geneontology.org/GO.downloads.annotations.shtml from GO Consortium members, as well as tools to access and process http://www.geneontology.org/GO.tools.shtml this data. This is a good example of a standardization initiative in an area where everyone was formerly using different terms for the same objects. This type of success requires cooperation among leaders in the field in question and a workflow that produces an accepted standard. Although not examples of standards, these papers, whose contributing authors include Rex Chisholm, Dean of Research for Northwestern's Feinberg School of Medicine, are examples of consortial projects dealing with identifying and re-using data:

1. Gaudet P, Bairoch A, Field D, Sansone SA, Taylor C, Attwood TK, et al. Towards BioDBcore: a community-defined information specification for biological databases. Database (Oxford). 2011;2011:baq027. PMID: 21205783. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3017395/

2. Gaudet P, Bairoch A, Field D, Sansone SA, Taylor C, Attwood TK, et al. Towards BioDBcore: a community-defined information specification for biological databases. Nucleic Acids Res. 2011;39(Database issue):D7-10. PMID: 21097465. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013734/

3. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011;3(79):79re1. PMID: 21508311.

4. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4:13. PMID: 21269473. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3038887/

Defining standards for data exchange is difficult, but a bare bones framework of required minimum fields for describing a dataset will be useful. Likewise, using tools like JHOVE2 or DROID to identify, validate, and extract features from data sets will greatly enhance compliance with description requirements by reducing the number of fields for which data must be manually supplied.

Thank you for this opportunity to comment.

Sincerely,

Daniel Linzer
Provost