Dr. Melissa Haendel, Ph.D.
Assistant Professor
haendel@ohsu.edu
Oregon Health & Science University
Portland, OR

On behalf of the Resource Discovery Group, a consortium of researchers from eagle-i (https://www.eagle-i.net/), Vivo (http://www.vivoweb.org/), the Neuroscience Information Framework (NIF; http://neuinfo.org/), Biositemaps, and the CTSAs, whom are interested in promoting research resource representation and discovery in the scientific enterprise.

Preservation, Discoverability, and Access

> *(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?*

Federal agencies should create a technical standard that enables discovery, usability, attribution, and long-term preservation of digital data. These specifications need at a minimum to include the archiving of data in publically accessible repositories, using standard record and metadata formats, and promoting best practices for interoperability and reuse, such as Semantic Web standards and Linked Open Data. Once the technical standards are there, policy can be established that requires data to be made available in a compliant manner as a deliverable of all federally funded grants and contracts, not only for those over $500,000. Grants with a data-sharing component should have a required budget line item for data sharing and archive.

A critical aspect of this policy will be to define “digital data” in the context of the policy. Funding agencies can support researcher efforts to meet the policy requirements by integrating semantic reference to these digital data into grant application and reporting structures. With appropriate tactical issues worked out, funding agencies could partner with publishers to require (and verify) data sharing before research results can be published. Finally, award and incentive systems (including institutional APT committees) must recognize the value of quality data management and sharing to the scientific enterprise.

It is estimated that it costs $24,100 and from 1.5 to 3 years to develop a transgenic mouse from scratch (eagle-i, unpublished economic analysis). What if that mouse were available to the research community at or during its development? This could expedite both public and private research endeavors. One of the issues is that “this” mouse is neither shared nor represented in a standardized manner such that it can be found for general reuse. It is not until a curator at a specialized database sees it in a publication, that it becomes part of the public record of available resources- sometimes years after it was developed. The point here is that the metadata about research resources themselves is digital data, and standardized representation and sharing of research resources should be included in any digital data policy. It should be noted that a lack of data annotation and sharing may not be for lack of desire to do so. Funding agencies, libraries, and research offices should offer training and helpdesk facilities to educate researchers in best practices for data annotation and sharing.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

One issue is that currently it is largely only publications and patents that are attributed. The scope of attributions needs to expand. Researchers need the ability to access the components within the publication (e.g. a knockout mouse, viral vector, database, datasets, etc.) This will protect the interests of individual stakeholders and they will feel more inclined to share these important and relevant outcomes of the scientific enterprise. Specifically, data sets can be citable, authored sets of information that can be referenced in the context of publications, grant reports, etc. While mechanisms are underway to support such efforts (Bioresource Research Impact Factor, Beyond-the-pdf, nanopublication), it will not be until funding agencies, employers, and publishers consider such citations in the context of evaluating a candidate proposal or manuscript that they will be adopted.

However, federally funded research produces data that is generated using taxpayer money and it belongs to the people. The person who generated it has no intellectual property interests on the *data.* What they do with it is a different matter, and they should be given some amount of time to do something with it (publish, patent, market, etc.)- 9 months or a year, perhaps.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

In developing policy that accommodates differences between scientific disciplines, libraries and information science researchers that are accustomed to providing guidance and resources for disparate kinds of data should be consulted. While data differs in different disciplines, there are qualities common to all data types, and these should inform inter-disciplinary requirements. For instance, there exist upper ontologies that represent the types of things that exist. Classification of data elements can be tied to such upper ontologies via reuse of these upper ontologies. One example is the Basic Formal Ontology as the upper level ontology for all Open Biomedical Ontologies (OBO; http://www.obofoundry.org), which enables representation of a catheter, a zebrafish liver, diabetes, and regulation of cell adhesion. These entities may not on the surface appear to have anything in common, but use of a common upper ontology can facilitate data integration about all of them (for example, in the context of designing an experiment). However, it is equally important to consult the end-user who is attempting to query across disciplines to ensure data consistency of representation. To this end, existing discipline/data specific repositories should also be consulted to ensure applicability. Furthermore, to support innovative reuse of digital data, it is important to recognize that these uses are not usually the original creator's intent. Data from disparate disciplines, projects and sources can be combined for synthetic and synergistic scientific inquiry - this in itself will also support new markets. Interoperability standards will benefit these new applications. Therefore, each discipline may require specialized data formats, queries and applications, but federal agencies can promote open and extensible standards to meet cross-disciplinary needs.

Another facet of this that must be considered is the extent to which there exist different data sensitivity issues in different fields. For example, publication about uranium enrichment metadata may require different consideration than data on the Arabidopsis genome.

> *(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?*

The most important aspect of what federal agencies can do with respect to garnering an understanding of the cost benefit analysis of stewardship and dissemination is to promote scientific inquiry that depends on public digital data. It is currently difficult to obtain funding for such projects, and as such, it remains somewhat of an idealistic rationale that publication and availability of data will be good for the research enterprise. In fact, we know that it is difficult to reuse others' data without standards, and it is often more cost-effective and time saving to create one's own data. If we are to tip the scales and actually save time and money, it will be because there exist standards and requirements to promote data reuse. Such requirements can be met via interagency collaboration, standardization, and cost sharing. In doing so, there is the potential to control costs and maximize benefits by limiting duplicate efforts, distributing responsibility, and by educating researchers. Furthermore, with respect to research resources, there is a clear indication that reuse of such entities saves time and money. If standards were promoted to enable their identification and relevance, and researchers incentivized via funding streams to leverage preexisting resources, this could lead to a very solid understanding of cost-benefit to sharing digital data for these particular data types.

> *(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?*

Participation by the many stakeholders must be regulated by technical and legal standards to ensure and promote free public access, discovery, re-use, and preservation. The expertise and methodologies of these stakeholders should be leveraged collaboratively both in the development of policy and in its execution. Such collaboration is required for success, and can drive best practices, innovation, market creation, and compliance.

The present repositories of research communities, publishers, and institutions can be utilized and developed (e.g. Pangea, TreeBase, eagle-i, NIF, Biositemaps). Existing partnerships between publishers and repositories, such as Dryad, can be grown. Organizations like DataCite and BioCoreDB work to improve the discoverability and utility of data. However, none of these systems alone will be successful until they are integrated into the research workflow. It has to become easy to submit data to such repositories in the context of publishing manuscripts or submitting grant reports. These repositories must also supply the submitters with some form of unique identifier. These identifiers can be used to track submissions and, eventually, resource usage.

Universities, research institutions, and libraries will need to play a key role in building infrastructure to support their researchers' compliancy and education, as with NIH public access policy, and guiding archival and discovery standards. They must also include data sharing and stewardship as a component of performance evaluation, where applicable.

Libraries are also well positioned to enable these infrastructures to be compliant with the Semantic Web and population of Linked Open Data from these data sources.

> *(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?*

The scientific and economic reward of sharing digital data will not be realized until the cost of data management and preservation are built in as part of any research program (whether it be in the context of a grant, laboratory management, a library, etc). Funding agencies should require researchers and institutions to document the cost of data management and publication within their proposals and reports. As this would be a new policy, better guidelines for what types of data management and preservation are satisfactory should be developed. Included in these guidelines would be requirements for sharing metadata about research resources. These guidelines should also promote the collaboration and/or inclusion of information specialists or libraries in supporting this aspect of the research. Furthermore, agencies should consider funding information scientists and libraries to perform more research on making specific data types conform to standards and archived for maximum query potential. In summary, information scientists now are needed more than ever to be a part of the research endeavor rather than solely involved in after-the-fact archival activities.

> *(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?*

Interagency standards that offer practical workflows and mandate deposit in publically accessible repositories will improve compliance and facilitate verification. These standards would require that: Digital data and research resource metadata are deposited in publically accessible databases in conjunction with manuscript acceptance and final grant reports, and that standardization of data format and minimum metadata are applied and *verified*. Different levels of compliance would need to be defined. At the very minimum, data must be understandable and reproducible based on free text descriptions and "readme" type files. Higher levels of compliance, e.g. structured metadata to enabled querying and reasoning across datasets, would be optional.

Many publishers already require certain data sharing standards and yet authors do not always comply in spirit or in letter. In support of these standards, it is recommended that several submission workflows be supported, including third-party deposit. One very important aspect of this is to involve the publishers, in particular with the assignment of persistent, unique, and linked identifiers. Currently, a manuscript may be published wherein the subject is a unique gene (or some other common data element), and yet these elements are never uniquely identified. There **must** be a partnership between researchers, publishers, reviewers, and funding agencies to ensure that such entities are properly referenced. Only then will their reference be linked to the research landscape and enable maximal inference and discoverability. Perhaps even more importantly, only then will the research be reproducible. This is especially relevant in the context of research resources, where without reference to a specific resource ID (for example, an antibody ID) one will never be able to reproduce the experiment let alone find the resource. For verification, publication of digital data on the Semantic Web can further enable systematic review of the data.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Federal agencies could establish incentives to stimulate the use of research data through preferential review of research proposals producing and leveraging public data in addition to requirements for archiving data. For example, NIH grant guidelines could be modified to include leveraging public data as part of a grant's Approach or Environment scores. Application showcases (e.g., http://www.data.gov/developers/showcase) or contests can also raise the profile of public data and capture the attention of the media on data standards and public availability of data. Small-scale venture capital solicitations patterned on http://www.kickstarter.com and data marketplaces such as http://www.crunchbase.com offer models for value-added services on top of data where relatively small investments of capital could produce significant results in the private sector.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

As with manuscript publication, secondary results should cite primary data. Standardized unique data identifiers will enable identification and linking resource (be it data, research resources, etc.) to relevant documents, data, persons, and grants. The use of controlled author and institutional identifiers (e.g. ORCID registry, http://orcid.org) will be critical to support disambiguated and resolvable attribution. Furthermore, use of a common metadata standard to tag various kinds of data with appropriate attribution in a standardized way will ensure proper attribution. It is not always enough to know whom the data came from, but also the version, from where, and how is it related to other documents, data, experiments and grants. Simply stating the author and the year is not sufficient to understand the methodology or process in which the data was reused. These additional metadata could promote a standard for provenance, quality and trust of scientific data.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

First and foremost, a minimum attribution standard for any kind of content should be created. Anything that is reportable as linked to grant funding activity should meet this standard. Following that, it will be important to develop metadata standards that facilitate machine reading and Semantic Web linking of information. Such metadata standards can be high level, as per the upper ontologies mentioned above. Basically, what kind of resource is it? Who, where and when is it attributed to? What is it linked to? Following this, each discipline will have further requirements and standards to better inform reuse in those fields. However, a simple adherence and strategy for including the aforementioned metadata will support and inspire more extensive data annotation. In the context of research resources, we are working on a metadata standard to this end. Publishers and granting agencies can adopt this metadata standard and provide guidance to contributors in support of meeting this new standard.

> *(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?*

There are successful standards developments in many domains. The W3C standards process has successfully produced HTML, XML, RDF and other languages. Key to the success has been its openness and community participation. Successful standards development relies on the contributions of a diverse population of experts, including scientists, information professionals, and technologists. It has to be field tested- if it is not useful or doesn't work for end users then it will not be adopted. If scientists themselves begin to reap the benefits of standardization, they will no longer feel the burden of having to comply. For example, if they can search all completed grants for specific research resources that may be advantageous to their work, and then find some that they reuse, they will not feel such a large obligation when it is their turn to provide the metadata necessary to make their own research resources available.

> *(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?*

A technical infrastructure that utilizes international standards for interoperability and re-use, such as Semantic Web Standards and Linked Data, should be adopted. Agencies can and should leverage the work of organizations focused on international data sharing and utility, such as CODATA, the Global Biodiversity Information Facility, the Open Archives Initiative, and the Digital Curation Center. It would also be worthwhile for federal agencies to participate in and support international efforts to connect data collections and build collaborative data infrastructures that aim to deliver cross-disciplinary data services. Similarly, adoption of other international efforts to standardize metadata, for example, coordination between VIVO and EuroCRIS, the European organization for international research information, will facilitate data integration internationally. Promoting such coordination as part of existing granting mechanisms or via new ones to promote international collaboration will be beneficial. Mechanisms could include specific RFAs for projects that coordinate internationally and hosting international workshops to bring these groups together.

> *(13) What policies, practices, and standards are needed to support linking between publications and associated data?*

To facilitate linking between publications and data, the use of persistent, unique identifiers for data, research resources, publications, authors, and institutions is required. These identifiers should be unique Internationalized Resource Identifier (IRI)—the standard for identifiers on the World Wide Web, so that the data (or the metadata about the resource) can be made directly available trough the web. Unique identifiers enable visible links between entities, as well as re-use and the development of new services. For example, browsing a publication could include integrated data displays. In support of this functionality, we need standards for citing datasets and models, along the lines of what SageCite is working towards. Further, linkouts between publications, datasets and resources are needed; similar to the way they work today for genes. Clicking on links to research resources could take you to a place where you can obtain the resource. Retrospective curation, at least for major datasets and publications (e.g. TCGA, Wellcome Trust, etc.) should be considered.

Disambiguation services such as the Virtual International Authority File (VIAF, http://www.oclc.org/research/activities/viaf/) offer a promising path forward for improving data quality. While VIAF focuses on organizations and people, other much lighter weight efforts could be established using open tools such as Google Refine (http://code.google.com/p/google-refine/) to support disambiguation web services from data repositories that could be integrated into desktop systems, websites, and publication submissions tools. Services that enable linking to data and linking both data and publications to known identifiers or terminology at the time of submission of a new publication could push much of the linking upstream to where incentives for documenting work are the highest.

Enabling such capabilities will require a new age of semantic awareness on part of the researcher, the reviewers and the publishers of manuscripts and data. Enhancing current research training to include modern information management strategies will be key, and funding agencies should support integration of information management into their research workflow.