

Data science holds the future of cutting-edge analyses

By Eddy Ball and Kelly Lenox

The growing priority of data science at NIEHS, to glean crucial insights from the vast amounts of information generated in biomedical research, was underscored by two September events. A two-day workshop Sept. 15-16 focused on progress toward a consistent language for environmental health research. It was followed by a Sept. 17 mini-symposium, which was co-sponsored by the NIEHS Data Science Seminar Series and Keystone Science Lecture Seminar Series.

“Workshop for the Development of a Framework for Environmental Health Science Language” brought together an interdisciplinary group of experts at North Carolina State University (NCSU) in Raleigh. Chaired by Carolyn Mattingly, Ph.D., of NCSU, and Melissa Haendel, Ph.D., of Oregon Health and Science University, the workshop explored a framework for creating standard languages to ensure that descriptions and content of data sets can be understood by the broader research community.

Consistency in language and terminology is a crucial step toward enhancing reproducibility, data reuse, and data integration. The mini-symposium, held at NIEHS, highlighted several areas of research that have applied data science techniques to advance diagnosis and treatment of disease.

Developing a common language

NIEHS and National Toxicology Program Director Linda Birnbaum, Ph.D., opened the workshop and introduced the major themes of the presentations that followed. “We really need to begin to integrate the huge amounts of data being generated,” Birnbaum told the attendees. “A common language is very important, and we’re going to have to be bold,” she said, referring to the collaboration and knowledge management that are essential to the task.

Along with representatives of universities, agencies, and other organizations, a number of NIEHS scientists and grantees participated in the meeting. After two keynote presentations, 11 short talks, a panel discussion, and two breakout sessions, participants spent the remaining half-day drafting reports that will lead to a paper outlining plans to advance the effort toward a common language in environmental health science data.

But, as Haendel cautioned early in the workshop, that’s just the first step. “Any ontology worth its salt is never done.” Still, the workshop itself was a major leap forward for a working consortium that had humble beginnings with a local meeting in the summer of 2013 (see [story](#)).

Data science advances underway



Boyles encouraged audience interaction with speakers at the mini-symposium, fostering an energetic exchange. (Photo courtesy of Steve McCaw)



According to McCray, the variability and severity of ASD symptoms suggest that genotypic links may help tease out underlying conditions. (Photo courtesy of Steve McCaw)

Becky Boyles, data scientist in the NIEHS Office of Scientific Information Management, opened the mini-symposium, which drew a full house. “It’s the first data science mini-symposium we have sponsored,” she said, “and we’re very open to what people want to see in future seminars.” Boyles was joined by Astrid Haugen, science program analyst with the NIEHS Division of Extramural Research and Training, who co-hosted the talks.

Three invited experts in data analysis focused on autism spectrum disorder (ASD), the microbiome, and model systems (see [text boxes](#)).

Alexa McCray, Ph.D., director for biomedical informatics at Harvard Medical School, presented her work modeling ASD. McCray and her team work with highly detailed phenotyping, or collecting clinical, biological, and other observable data. They have developed a set of consistent terms, or ontology, that enables researchers to integrate and query disparate ASD datasets.

Owen White, Ph.D., lead investigator for the data coordination and analysis center of the National Institutes of Health Human Microbiome Project

(<http://commonfund.nih.gov/hmp/index>)

, engaged the audience in an animated discussion of the importance of data science, by using the example of metagenomics, which is the effort to analyze the genomic data of humans and of the vast array of microbes involved in human metabolic functions

Melissa Haendel, Ph.D., assistant professor in the Oregon Health and Science University library, highlighted another data science challenge with great promise. The

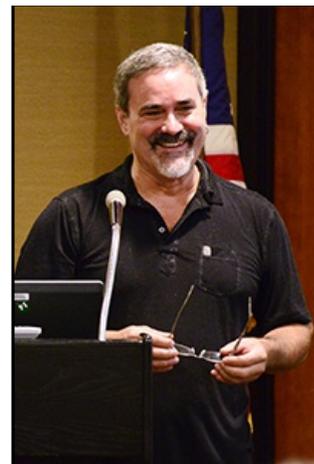
[Monarch Initiative](#)

(<http://monarchinitiative.org/>)

is a translational research project that provides tools for navigating great stores of genetic and phenotypic data across species, to advance diagnosis and treatment of diseases that are rare or previously undiagnosed.

Linked Video

[Watch as White explains the work at the Human Microbiome Project Data Analysis and Coordination Center in this University of Maryland video. \(0:32\)](#)



“Data integration, regardless of any other implementation, is achieved by creating controlled vocabularies and ontologies,” White said, emphasizing that it is a consortium-driven process. (Photo courtesy of Steve McCaw)



According to Haendel, the benefits of data science advancements include searchable details on author contributions and useful details on resources — for example, a link to the genome of the mouse used in a study and links to other studies using the same strain. (Photo courtesy of Steve McCaw)

McCray — Data access to improve assessment of ASD

“An ontology is a way to characterize the world, taking concepts, as a thesaurus does, in order to understand relationships among them,” McCray told the audience. “An ontology provides a structured, computable way of representing a domain of interest, such as ASD.”

In addition to collecting biological samples for genetic data, researchers used a battery of scoring instruments on 500 families affected by ASD, including interviews, questionnaires, and direct clinical assessments. In the ontology, published in a 2013 paper, McCray’s team developed various concept categories for different types of data. For example, the ontology’s category of self-injurious behavior associates the frequency of head banging measured in one instrument with the frequency children hit themselves against an object, which was measured in another instrument.

According to McCray, use of the ontology, which is publicly available, will enable phenotypic data to be correlated with genome-wide association studies, in the hopes of finding genetic links for this highly heritable condition. Discovery of genotypes related to autism could lead to earlier diagnoses, enabling early interventions, which often have marked results.

Citation: [McCray AT, Trevett P, Frost HR.](#)

(<http://www.ncbi.nlm.nih.gov/pubmed/24163114>)

2014. Modeling the autism spectrum disorder phenotype. *Neuroinformatics*. 12(2):291-305.

White — Microbiomes, metabolic pathways, and metagenomics

“There is more metabolism going on in your microbiome than in any other organ of your body, so it really functions as another organ,” White said, opening his presentation. However, research may be hampered by databases that do not conform to a larger ontology or control vocabulary, or that omit metadata, such as whether a sample was collected from an individual who was healthy or ill. These shortcomings limit the usefulness of data queries by scientists, who might otherwise find information to further their research.

“Data integration is achieved by creating controlled vocabularies and ontologies,” White said, referring to the importance of using key words people agree on, which he praised as a process of high scholarship.

White underscored the growing importance of data coordination centers, or curators, who regularize the data so it can be integrated. This will also allow scientists to compare studies that use different data sources.

Haendel — Using model systems to compare undiagnosed diseases

According to Haendel, numerous databases include disease and phenotype information, but it is difficult to make connections across the sources. Clinical data and the results of studies on model organisms, such as fish and mice, are available, but, as McCray noted with autism studies, phenotypic data are described in different terms across the sources.

“Constipation, in humans, is described as decreased gut peristalsis in zebrafish,” Haendel noted. “Standardizing phenotypes is the final frontier and will enable us to compare phenotypic features of different species, so we can understand the underlying genotypes and environmental causes that give rise to them,” she said.

Haendel and her colleagues built an ontology that connects anatomy ontologies to genotypes, incorporating the different genome data of each source. The resulting ontology, dubbed the [Uberon](#)

(<http://uberon.github.io/>)

, makes it possible to query all organisms in the system for a particular phenotype and return genotype data associated with it.

Haendel noted that integration of environmental exposure data is lagging further behind, due to the lack of standardized language — a challenge that the previous two-day workshop, which she co-chaired, helped move forward.

Citation: [Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA.](#)

(<http://www.ncbi.nlm.nih.gov/pubmed/22293552>)

2012. Uberon, an integrative multi-species anatomy ontology. *Genome Biol*. 13(1):R5

, Office of Communications and Public Liaison. The content is not copyrighted, and it can be reprinted without permission. If you use parts of Environmental Factor in your publication, we ask that you provide us with a copy for our records. We welcome your [comments and suggestions](#). (*bruskec@niehs.nih.gov*)

This page URL: NIEHS website: <http://www.niehs.nih.gov/>
Email the Web Manager at webmanager@niehs.nih.gov