Office of Data Science

Presenter: Ms. Stephanie Holmgren, NIEHS/DNTP/Office of Data Science
Dr. Charles Schmitt, KGS

The success of data-driven methods in the private and public sectors is driving new opportunities, many of direct relevance to Environmental Health Sciences. A large number of ongoing and new projects are providing new life science data sets in areas including genomics, metabolomics, and exposomics (e.g., TCGA, CHEAR, ECHO), a trend that will continue as new data collection technologies such as tissue chips become common. Likewise, access to sensitive data, especially patient and socioeconomic data, is increasing due to collaborative efforts such as the PCORI network, i2b2/Shrine network, BD2K MD2K, among others. Specifically within the toxicology community, high throughput approaches, biosensors, and predictive toxicology methods are driving the need for toxicoinformatic tools to manage and analyze the resulting complex and large-scale datasets.

Despite these opportunities, real challenges exist in making use of the growing wealth of data and tools. NIH and other agencies are now promoting FAIR data practices to ensure valuable research data sets are findable, accessible, interoperable, and reproducible. We introduce the notion here of FAIR+, in recognition that data must also be computable, a challenge that is increasing due to the growing volume of data and increasing security and privacy concerns.

For Environmental Health Sciences, these challenges are of special importance. The sheer number of environmental factors and the combinatorial complexity of interactions between factors and biological systems argues for data collection and curation efforts that are of high efficiency and quality.

The Office of Data Science was recently established to support a holistic view of data science at NIEHS to leverage the opportunities and address the challenges of data-driven research. Towards that end, the mission of ODS is to advance scientific discovery, policy development, and decision-making within NIEHS by developing and applying policies, standards, best practices, and technologies that increase the discoverability, accessibility, and utility of data.

The office's primary responsibilities address the following five pillars that support data-driven scientific discovery and the ideals of FAIR+.

1. ***Governance*** is critical to ensure policies and practices that promote preservation, access, interoperability, and reproducibility are in place while still preserving the rights of research subjects and the rights of data collectors to benefit from their efforts.

2. ***Data cyberinfrastructures*** need to be provided that allow researchers to discover, access, integrate, compute on, and share data and tools and that enforce governance policies.

3. New ***methods*** are needed, not only to analyze data, but also to improve the capabilities and cost-efficiency for how we store, represent, integrate, process, and manage data sets. The success of data science in a variety of fields (e.g., finance, marketing, pharmaceuticals) is enabling a rapid growth in such methods, but ***translation*** of these methods to Environmental Health Research is needed.

4. ***Training*** on these methods is critical, both to scientists who may be adopting the methods, but also for the IT staff, data and computer scientists, and informaticists who are building tools and systems to support the methods.

5. An active and collaborative ***community of practice*** within NIEHS is needed for those focused on the challenges to enable FAIR+ for environmental health research and to advance data- and knowledge- driven methods.

The office will be or is currently engaged in several activities that align with each of these pillars. With respect to cyberinfrastructure, the office is leading development of a Data Commons, a data management platform that enables NIEHS researchers and core labs to access, find, and share research data and metadata. This Commons will eventually interface with other NIEHS (CEBS, ICE) as well as external systems. To address the challenge of findability and interoperability, the office is designing a metadata service catalog to promote usage of standardized vocabularies for various applications. Finally, a collaborative NTP/NICEATM/EPA project is underway to use natural language processing (NLP) approaches to automate steps in the systematic review process. These projects are an example of several activities that the Office of Data Science is undertaking in order to realize the full potential of environmental health data to promote human well-being.