DataScope: An Information Instrument for the Future of Science

Katie Knight and Arjun Shankar, ORNL

The Facility Vision

We propose a suite of services that provide access to data that is *semantically connected and searchable*. This is a one-stop shop to help researchers find what data is available, understand how to get that data, clarify under what conditions that data may be used, and how it might be integrated with other datasets. These services are a means to increase the interconnectedness of data, and thus increase the value of knowledge. The facility provides access to the data that produced those publications, as well as the ideas put forth within the publications. *AI models will continuously create "nanopublications", and prioritized linked data sets, where basic ideas within papers are turned into snippets, giving an automatic context to the dataset or groups of datasets*.

Institutional Incentive

Large-scale data sources are the norm in most disciplines, and enormous data sets are produced on a regular basis by research groups across laboratories at an ever-increasing scale. For instance, the National Human Genome Research Institute produced 150 billion base pairs in 2007 (National Research Council, 2010), a number that now might be a day's output by one biology research laboratory. This rate of data generation outstrips the ability of a single investigator to glean insight from it and instead requires the analytical power of an entire community. But, this is not possible without data that is easily discoverable [MIB1]and analyzed. *We envision a future facility that serves as a data exploration instrument or a "Datascope".* Repositories of simulation and observational data, Datascopes can autonomously explore multi-modal scientific data for patterns, predict trends, make connections across domains, and serve as an accumulative store of the "sum of all knowledge". This knowledge is preserved in the bits of output, bits that represent the theory, models, and context to create data, and bits that capture value-added links and expansive metadata.

Scale and Scope

We attack the problem of data fragmentation across individual research landscapes seen in varied sampling methodologies and notation methods, data formats, domain specific regulations, creation tools, funding models, and concierge data expertise. We need availability, integration, searchability, and reusability. Our data must sing: within *and* across disciplines!

Imagine the paradigm of the "publication as center of scientific knowledge" inverted. As the Copernican revolution introduced the Heliocentric model, we propose a Datacentric model, where data (and theory and models represented as data) are positioned at the heart of scientific discovery. In this discovery platform/suite of services, the dataset sits at the center, orbited by publications produced by this dataset, software that generated the dataset, and documentation that gives context to the data itself. Part of the discovery process is apprising the searcher of both what knowledge has been produced from the data and how the data came to be, and potentially inspiring deeper or appurtenant research. This system will provide the tools, the data, the infrastructure, and the metadata to make felicitous and unexpected discoveries possible.