SEAD: Finding the Long Tail lost in Big Data
Margaret Hedstrom and Jim Myers, University of Michigan

White paper for the formation of a National Data Service Consortium
Boulder, Colorado, June 12-13, 2014

The "long tail" of data is constituted of data, users, research questions, and methodologies that differ in some fundamental ways from the "head" of the distribution, commonly known as big science. Unlike big science where research depends on a relatively small number of very large data sets that are well known or discoverable and "big" in the sense of volume, coverage and user base, scientists in the long tail produce and use massive amounts of data that is widely dispersed, maintained in millions of files that are difficult to discover and use, and valuable because of their unique coverage, times span, observations, variables, etc. In SEAD, we contend that social, economic and to a lesser degree, technological mechanisms, for data sharing, reuse, curation, and ultimately innovation in the long tail differ considerably from those that we associate with Big Data today.

In SEAD (Sustainable Environment/Actionable Data), a NSF DataNet project, we are working with sustainability scientists to develop cyberinfrastructure and sustainable services for data access and preservation in a community that is typical of the long tail. Some characteristics of this community include:

- Focused on problems that require data, methods, tools, and expertise from multiple disciplines
- Requires many different types of data about physical, natural, and social phenomena in order to understand interactions between natural and human systems
- Uses a combination of observational (field) data, experimental data, simulations, and models
- Conducts research in small to medium-sized labs or centers under the direction of a single PI or a Center Director.

Research communities, like the sustainability science community, exhibit several phenomena that are typical of the long tail, such as:

- Data discovery is via targeted foraging and word-of-mouth
- Almost all data are stored locally
- Metadata standards and ontologies, where they do exist, are based on disciplinary norms or local practices
- Data formats and metadata standards are often controlled by multiple independent third-parties (e.g. instrument and application providers, larger data providers) and must be translated/integrated as part of the research project.
- Researchers know how to work with data that is "good enough" but not perfect

- Data are vulnerable to interruptions in organizational arrangements (graduate students finish PhD's and move on – lab or center funding sunsets)
- No single data set is likely to have great value standing alone, but aggregated, combined and integrated data become valuable resources of discovery and innovation.

These characteristics of the long tail suggest that some of the strategies used to guide investments in infrastructure for big data (investing in a small number of heavily used and highly curated data resources, developing standard metadata schema, data representations, and file types, and setting priorities based on disciplinary demands) many not work well for the long tail. These types of investments are valid and much needed for scientific research infrastructure, but the long tail may benefit more from new approaches to infrastructure development tuned to a different set of social and economic characteristics.

One barrier to data reuse is the difficulty of discovering data that might be valuable for a particular study, model, or decision. Making data minimally discoverable saves the community time expended on futile searches and creates a market, of sorts, for the data. Researchers are notoriously disinclined to invest their own time and effort on data management when there is no indication that they or anyone else will benefit from their efforts. Creating very low barriers to entry to a network where data can be discovered and acted upon vastly reduces this disincentive to sharing data. Enabling researchers and their collaborators to make small incremental improvements to data as they use them makes data curation an integral part of sharing and using data.

In SEAD, we are developing such mechanisms for the sustainability science community. Researchers in the long tail of science can use the SEAD Active Content Repository (ACR) to drag and drop files and documents for sharing among members of a project, preview data, add comments and metadata, and add value to data while they are actively using it for research. We are building a research network using Vivo to link data with publications with people (authors, data producers, etc.), support visualizations of citation and co-citation networks, research contributions over time, and diversity of disciplines that are engaged in sustainability science. The SEAD Virtual Archive accepts data from the ACR or directly from research projects, packages the data with additional metadata, indexes data sets, and transfers the data to an appropriate repository for long-term preservation.

One success metric for SEAD's approach would be the emergence of network effects where locally stored and heterogeneous data sets become more valuable because they are discoverable, easily accessible, and shared among networks of researchers with common goals or similar research interests (be they scientists, policy makers, educators, students, or citizens). In the long tail, this number of users is unlikely to be very large for any particular data set. But, if the costs of creating an environment where this type of interaction can take place is low and widely distributed across

many different data producers and consumers, a small increase in data usage can have a large impact. This approach has the additional advantage of low organizational costs because it leverages the self-organizing behaviors of emergent communities.

We believe that any National Data Service must attend to the needs of the long tail of science that is underserved by and cannot take advantage of infrastructure that requires large local investments, technical expertise and support, and conformance with complex data and metadata standards.

Acknowledgements: SEAD Co-PIs: Praveen Kumar, University of Illinois Urbana-Champaign; Beth Plale, Indiana University, and the SEAD team: www.sead-data.net.

NSF Cooperative Agreement: #ACI-0940824