Professor Beth Plale
Director, Data To Insight Center, Indiana University
Science Director, Pervasive Technology Institute, Indiana University
Research Data Alliance/US Steering Committee
Research Data Alliance (RDA) Technical Advisory Board Co-chair
Co-PI Sustainable Environments Actionable Data (SEAD)

As the great database researcher Jim Gray said shortly before being lost at sea, "we have to do better at producing tools to support the whole research cycle—from data capture and data curation to data analysis and data visualization". That was in 2007.

Seven years later we have seen considerable research and innovation in workflow systems that make it easier to carry out analysis on data by separating out the orchestration of tasks from the nature of the tasks themselves. The Kepler, Swift, and Pegasus workflow tools have seen good use in the US. Workflow systems underpin today's Science Gateways which carry a non-trivial portion of the XSEDE use. The LEAD (Linked Environments Atmospheric Discovery) Science Gateway was one of the first science gateways (Gannon et al. 2007).

Machine learning and deep learning have exploded in computer science over the last 5 years; applicants for new faculty positions in these areas abound and this is surely to the benefit of the geosciences.

But data capture, data curation, and data sharing have lagged behind. Data mining and analysis require clean and tagged datasets. There are very few of these, and fewer, to my knowledge, in the geosciences. ***This workshop could promote the development of geoscience relevant cleaned and tagged datasets.***

The argument for data sharing is the reuse of data collected for one scientific or scholarly purpose by another researcher who may be from the same or different discipline. The data in a reuse case is used to verify the same result, to combine with another data set and answer similar questions or to answer a different research question. For instance, when weather data collected over decades is used for crop forecasting or water conservation. There are few success stories in data reuse in the geosciences***. This workshop could promote development of reuse success stories that use data collected in one sub-discipline of geosciences for use in a second sub-discipline. The major repositories are now part of a consortium. What could be learned by a reuse case that crosses two to three of these repositories?***

Everyone talks about data provenance but we don't have anything working yet. Not on a wide scale. ***Yet*** data provenance captured from scientific applications is a critical precursor to data sharing and reuse. ***Can this workshop bring together like minded people to address the problem of lack of stable provenance such as through a notion like "trust threads" that weave delicate connections between data sets used in science and technology.*** These connections, fine and minimal as they are, say something about a data set. The data set is able to prove something about itself, to convey its family name and lineage so to speak, and in doing so contributes to its own trustworthiness.

Our perspectives are drawn from a 5-year project funded by the National Science Foundation called Sustainable Environments Actionable Data (SEAD). SEAD is responsive to the expressed needs of sustainability science researchers for long-term management of heterogeneous data by developing new capabilities for data integration, dissemination, and long-term preservation. SEAD provides researchers with tools for active curation and uses social networking to engage data producers and data consumers in community curation, gradually shifting curatorial and collection development responsibilities closer to the point in time during which data are first created.

SEAD focus is on the "long tail" of social and environmental data: derived data products, data collections from individual researchers and small groups, and data sets of local, regional or topical significance that are critical to sustainability science but are of limited value until they can be integrated and referenced geo-spatially and temporally, combined with related data and observations, and modeled consistently. While most of the individual data "long-tail" data sets are small compared to the commonly discussed big data sets, their changeability and heterogeneity pose the same challenges as other "bigger" data.

References

Cheah, You-Wei and Beth Plale, Provenance Quality Assessment Methodology and Framework, *ACM Journal of Data and Information Quality*, *Special issue on Provenance, Data and Information Quality, Vol 5(3), Dec 2014.*

Chen, Peng, Beth Plale, Mehmet Aktas (2013). Temporal Representation for Mining Scientific Data Provenance, *Future Generation of Computer Systems*, Elsevier, Vol 36, pp. 363-378

Gannon, Dennis, Beth Plale, Marcus Christie, Yi Huang, Scott Jensen, Ning Liu, Suresh Marru, Sangmi Lee Pallickara, Srinath Perera, Saotshi Shirasuna, Yiming Simmhan, Alex Slominski, Yiming Sun, Nithya Vijayakumar 2007. Building Grid Portals for e-Science:

Gray, J. (2009). eScience: A Transformed Scientific Method. In Hey, T., Tansley, S., & Tolle, K. *The Fourth Paradigm: Data Intensive Scientific Discovery*. pp. xix – xxxiii. Seattle: Microsoft Research.

Plale, Beth, Eran Chinthaka Withana, Chathura Herath, Kavitha Chandrasekar, and Yuan Luo 2012. Effectiveness of Hybrid Workflow Systems for Computational Science, *Int'l Conf on Computational Science (ICCS), Procedia Computer Science, Elsevier,* Vol 9, pp. 508-517