Earth system science as a data-intensive knowledge domain, and the relevance of a Linked Data methodological framework to facilitate its data handling functions

Bhaskar Ramachandran, Terrestrial Information Systems Lab., NASA Goddard Space Flight Center, Science Systems and Applications, Inc., Greenbelt, MD 20771

1. *What are the key challenges or questions for Earth System Science across the spectrum of basic research, applied research, applications, and/or operations in the coming decade?*

Earth System Science (ESS) embodies a holistic approach to understanding the geophysical, biophysical, and biogeochemical processes, interactions, and exchanges that sustain all life-support systems on Earth. In the last few decades, the attempts to deconstruct and understand the changing states of Earth's life-support systems have led to a realization that studying natural and anthropogenic processes in isolation is not conducive to revealing the complex nature of interactions and feedbacks that are better exemplified as emergent properties through their coupled systems behavior. This realization not only justifies an ESS approach, but also propels a key challenge to sustain ESS as a data-intensive knowledge domain. Consequently, we should distinguish between approaches that address the need to focus on improving and understanding the domain science knowledge versus those that address smarter ways of handling the data that contribute to that domain knowledge. This paper's thesis advocates the latter as a viable approach to articulate and better understand ESS epistemology, and develop intelligent ways to handle the data and information to delineate complex environmental change dynamics.

ESS is both data- and computing-intensive. Therefore, conducting either basic or applied research and applications entail the acquisition, compilation, processing, and analysis of data from varied knowledge domains. Besides multidisciplinary inputs this may require deep knowledge integration through transdisciplinary sources (socio-economic, demographic, etc.) as well. Hence, this involves data sources with disparate methodologies, ambiguous terminologies, and a motley mix of data and metadata formats and standards that could potentially create bottlenecks for data handling, representation, discovery, integration and interoperability. Further compounding this are increasingly massive volumes of data that vary across types, qualities, and pedigrees, which necessitates smart data management solutions.

Technology offers an approach to meet the challenge of dealing with this burgeoning data- and computing-intensive science. Since the current incarnation of the World-Wide Web was created in 1989, the same technologies are continuing to help create machine-encoded semantics that can potentially transform how humans and machines interact with what is called *Linked Data*, a method to create and publish structured data based on current Web standards, and a set of best practices. Linked Data are semantically enabled and machine-readable, and hence, shareable, queryable, and reusable across applications, and offer an efficient conceptual model for intelligently handling and managing ESS data. The term *Linked Data* is an integral of the original concept of *Semantic Web*, which itself has transformed from its original design. It can facilitate ESS by providing a formal mechanism to define, represent, and integrate its concepts and relationships based on contextual semantics to discover patterns, relationships, and predictive analytics as well. This technology has the potential to leverage a paradigm shift, but a major challenge for ESS is consensus acceptance. When compared to early adopters like the medical and bioinformatics[1] communities, the Earth science community has yet to invest in these technologies in a significant fashion.

Data-intensive science poses a serious set of challenges to data management and stewardship for ESS in the next decade; we have already experienced symptoms of these thus far. They relate to strengthening attribution, reproducibility, and provenance for ESS research and applications. All three areas will help alleviate skepticism from those who do not understand, misunderstand, or consciously obfuscate the science of global environmental change by stymying the progress to accomplish objective policy-based science[2]. Linked Data facilitates addressing all three challenges.

The final challenge is more sociological. The epistemic culture[3] of different knowledge domains determines how successfully they encourage (or discourage) their practitioners to use new methods and approaches. The epistemic culture within ESS is a valuable source of information that is worthy of investigation to help reveal how idiosyncrasies of team

---

[1] Good examples in biomedical informatics and translational medicine include the following, respectively: Shortliffe, E., and Cimino, J. (Eds.) (2014). *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. Springer, NY., and Kashyap, V., et al. (2008). *The Semantic Web: Semantics for Data and Services on the Web*. Springer-Verlag, Berlin.

[2] Good examples of such obfuscation are described in Oreskes, N., and Conway, E.M. (2011). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press, NY, and Hamilton, C. (2010). *Requiem for a Species: Why We Resist the Truth about Climate Change*.

[2] Good examples of such obfuscation are described in Oreskes, N., and Conway, E.M. (2011). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press, NY, and Hamilton, C. (2010). *Requiem for a Species: Why We Resist the Truth about Climate Change*. Routledge, London.

[3] Karin Knorr Cetina defines epistemic cultures *as cultures that create and warrant knowledge*.

dynamics may induce cooperative, collaborative, or isolative behaviors to efficiently address the scope of research and applications over the next decade. Related to epistemic culture are recent developments in the Science of Team Science[4] that explore the attributes deemed necessary to achieve a collaborative team's scientific and translational goals, and also highlight the existing sociology of scientific knowledge.

2. *Why are these challenge/questions timely to address now especially with respect to readiness?*

The urgency of the Anthropocene is increasingly being manifested via humans and their activities that have become *a new global forcing agent*,[5] and *a global geological force in its own right*.[6] The increasing impact of human dominance is borne by Earth's different biophysical thresholds. According to some researchers[7] humans have already transgressed three of the nine planetary boundaries with very serious consequences for all life globally in the near future. Science and society in the next few decades must contend with how best to understand and interpret the dynamics of global environmental change, and help inform policymakers by carefully addressing different scales of governance vis-à-vis scales of analyses. Linked Data offers an informatics-based approach with an unprecedented capability in linking datasets representing natural and human processes to support an evolving knowledge ecosystem.

An array of national and international studies/programs[8] continue to deliberate on various aspects of global environmental change even as policy and governance aspects to understand, manage, and mitigate its impacts are often mired in ideological and geopolitical squabbles. Most studies concur with the seriousness of what is happening but do not

---

[4] National Research Council. (2015). *Enhancing the Effectiveness of Team Science*. Committee on the Science of Team Science, N.J. Cooke and M.L. Hilton, Editors. Board on Behavioral, Cognitive, and Sensory Sciences, Washington, DC: The National Academies Press.
[5] Zalasiewicz, J., et al. (2010). The New World of the Anthropocene. *Environmental Science & Technology*, 44, 2228–2231.
[6] Steffen, W., et al. (2011). The Anthropocene: conceptual and historical perspectives. *Philosophical Transactions of the Royal Society*, A, 369, 842–867.
[7] Rockström, J., and 28 co-authors, (2009). Planetary Boundaries: Exploring the safe operating space for humanity. *Ecology and Society*, 14 (2): 32.
[8] A sample would include the Intergovernmental Panel on Climate Change, the Millennium Ecosystem Assessment, the International Council for Science's Earth System Science Partnership, which includes the International Geosphere–Biosphere Programme, World Climate Research Programme, DIVERSITAS, and International Human Dimensions Programme, United States Global Change Research Program, Global Earth Observation System of Systems, European Space Agency's Living Planet Programme: Scientific Achievements and Future Challenges – Scientific Context of the Earth Observation Science Strategy for ESA, The Belmont Forum, etc.

approach or advocate intelligent methods to address the data-information-knowledge continuum as a priority.

More data are constantly being acquired and processed than the ability to analyze and distill knowledge from them in a timely fashion. Besides, the coupled use of data from multiple sources in conjunction with model simulations is only likely to increase in our quest to understand global Earth processes at various spatial and temporal scales. Therefore, investing in contemporary data handling technologies is imperative. As a complement to Linked Data, smart solutions also exist to build an efficient cyberinfrastructure as collaborative computational environments that provide hardware and software platforms, tools, services, networks, protocols, data systems, and people interacting within a distributed architecture. But ESS has yet to fully benefit from all these technologies.

3. *Why are space-based observations fundamental to addressing these challenges/questions?*

Remote sensing-based observations remain absolutely central to our ability to conduct ESS research and applications essentially because they provide panoptic views and time-series measurements of all elements of Earth's life-support systems that are impossible to replicate using ground-based methods. Alongside our space-based observation capabilities, we also need to invest in infrastructure that enables the machine-readability of the derived datasets to create ESS Linked Data to support derived knowledge management functions.

Given the budgetary constraints to fund future Earth observing missions, NASA should consider not only building partnerships with the commercial sector but also with friendly global allies who have reliable and established civilian space programs. The nature of the global environmental challenges in the near future may, of necessity, warrant such partnerships.