Big Data and HPC: Exploring Role of Research Data Alliance (RDA) Report On Supercomputing 2013 Birds of a Feather November 20, 2013 Denver, Colorado, USA

By: Beth Plale, Indiana University

The ubiquity of today's data is not just transforming what is, it is transforming what will be laying the groundwork to drive new innovation. Today, research questions are addressed by complex models, by large data analysis tasks, and by sophisticated data visualization techniques, all requiring data. To address the growing global need for data infrastructure, the Research Data Alliance (RDA) was launched in FY13 as an international community-driven organization. We propose to bring together members of RDA with the HPC community to create a shared conversation around the utility of RDA for data-driven challenges in HPC.

Mark Parsons, Secretary General of RDA, started off the event with the powerful point that all of society's grand challenges require diverse (often large) data to be shared and integrated across cultures, scales, and technologies. But how does global interoperable infrastructure emerge? According to Edwards et al. (2007), "Understanding Infrastructure: Dynamics, Tensions, and Design", a source that Mark finds extremely instructive here, the dynamics of infrastructure evolve over time:

- First the infrastructures become "ubiquitous, accessible, reliable, and transparent" as they mature.
- There is a staged evolution:
    - "system-building, characterized by the deliberate and successful design of technology-based services."
    - "technology transfer across domains and locations results in variations on the original design, as well as the emergence of competing systems."
- Finally, "a process of consolidation characterized by gateways that allow dissimilar systems to be linked into networks."

The Research Data Alliance addresses the four axes of the infrastructure ecology, shown in Figure 1. It encompasses both the local and the global, with Plenaries and working group and interest group activity being globally inclusive, and adoption and implementation of technology and processes done on a local scale. The organization by virtue of its inclusiveness has attractive activity to overcome challenges that are both technical and social. RDA's create-adopt-use philosophy and harvestable efforts within a rapid 12-18 month timeframe means that results can be achieved quickly.

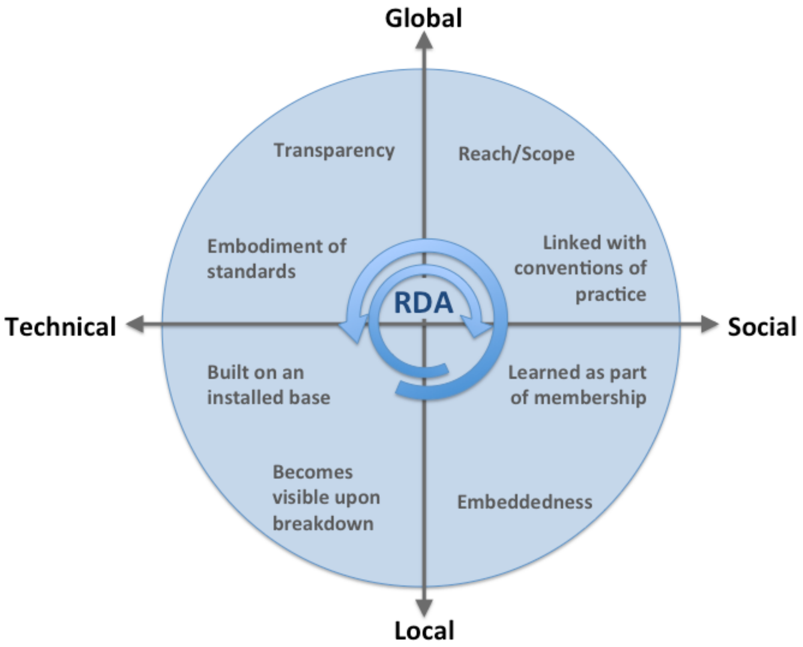
Global
Transparency
Reach/Scope
Embodiment of standards
Linked with conventions of practice
RDA
Technical
Social
Built on an installed base
Learned as part of membership
Becomes visible upon breakdown
Embeddedness
Local

Figure 1. RDA in the ecology of infrastructure. Derived from F. Millerand based on S. L. Star & K. Ruhleder (1996)

A panel took on the topic of RDA's role and relevance in high performance computing. Moderated by Mark Parsons, the panel undertook a discussion of the data challenges facing high performance computing, and the ways in which the challenges could be addressed by utilizing the global/local and social/technical strengths of RDA.

Panel members included:

- Nancy Wilkins-Diehr, San Diego Supercomputer Center, San Diego, CA; XSEDE project
- Fran Berman, Rennselear Polytechnic University, Troy NY; Research Data Alliance
- John Cobb, Oak Ridge National Lab; DataOne project
- Scott Lathrop, NCSA, University of Illinois; XSEDE project
- Beth Plale, Indiana University; Research Data Alliance
- Rob Pennington, NCSA University of Illinois; data architect

The panel members and audience engaged in discussion of several questions:

- What are challenges faced by big data/HPC centers in supporting applications that require data from international sources? Similar for data federation services (DataOne), and gateway services (XSEDE)?
- How are data/HPC centers dealing with data analytics applications, particularly those who have need for persistent large data sets?
- What are challenges and actions in HPC in response to recent attention on science reproducibility?
- Are broad data issues like reproducibility and data citation part of HPC's education and workforce development plan? What data issues are a part?

- Where to you see RDA contributing to these or other data challenges facing HPC?

Discussion was broad and interesting. The XSEDE project has a strong and long-standing focus on education, led by people like Scott Lathrop. The importance of early career engagement and opportunities was noted as important common goal in both XSEDE and RDA, articulated by both Scott Lathrop on the HPC side, and Beth Plale of RDA who has considerable experience in data science. There has been follow up between RDA and XSEDE on the early career aspects after the BOF. As Robert Pennington pointed out, HPC often has extremely large data sets, and established processes and standards for dealing with the data. Some of HPC disciplines, such as astronomy, have well-developed standards, and are global in their agreement. Long tail sciences are not so well organized, and an organization like RDA can provide a vehicle where within-discipline and across-discipline sharing and interoperability can take place. Nancy Wilkins Diehr, with her extensive experience in science gateways in Teragrid and XSEDE, had numerous good suggestions for advancements in data that would help the science gateway scientist. An obvious one is provenance, which is currently being addressed in an RDA interest group “Research Data Provenance”. A simple step forward for HPC, it was noted by a well regarded member of the HPC community, would be defining the minimal metadata required for every file deposited in a large tape archive.

Science reproducibility was raised as an issue. Where precise “repeatability” of a large model run or experimental evaluation may not be possible, reproducibility is possible through tools like data provenance and self-documenting services. Reproducibility and data citation should be critical parts of both HPC and RDA education and workforce development plans.

The BOF had 40-45 people in attendance.