The Science Data Highway - where are we today?

Silvia Nittel
University of California, Los Angeles
Computer Science Department
Los Angeles, CA 90025, USA
silvia@cs.ucla.edu

Abstract

The goal of this panel is to address issues related to the "Science Data Highway". The panel focuses on enabling technology, interoperability issues and problems, and outlines possible approaches as well as identifies inherent difficulties and limitations that might not be of technological nature.

1. Panel

Science data is distinctively different from other types of data: it is very large in volume, it has complex structure, it is not always complete and may have measurement errors in it, and people want to perform scientific analysis based on it. Therefore, they often want the data in a raw form, and if it is processed they are interested in the processing that has been applied to the data. For science data, several types of desired data interoperability exist: a) scientists are interested in retrieving data from centralized data providers which are often government agencies that operate e.g. space-stationed satellites or earth-stationed sensors, and b) they want to share data with each other. Traditionally, scientific computing required supercomputers, and scientists have been collaborating in close proximity at laboratories associated with supercomputer centers. With the advent of cheap and powerful personal computers and workstations, cheap storage as well as ubiquitous networks, scientists have been able to work spatially distributed today. Collaboration became a different face. Now, scientists are interested in data interoperability, because they need to share data over the network, over different computer platforms, and from different sources. The distributed infrastructure also opened up the possibility to collaborate with teams to solve larger problems.

What is data interoperability? Intuitively, it can be defined as "My applications can work with your data, and your applications can work with my data". In order to achieve this, we need to agree on several points: how to represent the data syntactically? Which meta data is allowed? How can we represent and agree on the semantic aspects of the data? Since we are also interested in online data access and exchange, we need to agree on a network communication protocol and infrastructure, a way to name and identify data, and a data server interface that allows us to retrieve data, possibly to preprocess and subset the data, before it is retrieved over the network. This also applies for science data interoperability.

Today, science data interoperability is quite in its infancy, especially for large sensor-based data sets such as in the Earth and Space sciences. This is understandable since the problem is comparatively new; however, it is also time to analyze the problem more deeply. Are there inherent difficulties of science data interoperability? What should the role and contribution of computer science and technological advances be to the problem? Why is the problem hard, and why are we not farther ahead yet? Is it all a technology problem, or which other factors play a significant role? Which type of interoperability do we really need for the science domain?

Having identified the goals and scope of science data interoperability, it can become clearer which approaches are more likely to solve the problem. Technology plays a significant factor, and so do standards. In this panel, positions will be taken on their roles, on possible approaches, and we will also identify problems that remain to be solved. Furthermore, the panel will include the representation of three different large-scale approaches for science data interoperability (NOAA, NASA, and Unidata) to elucidate the real world requirements, challenges, solutions, and open problems.