Shipping Computation to Very Large Data – an Old Idea Demanding Reexamination

Clifford Lynch
Coalition for Networked Information
Cliff@cni.org

February 28, 2020

Shipping computations to data rather than transferring a copy of data to a computational resource is an old idea. This can be seen in a variety of implementations dating back to the 1980s ; three notable examples are the "knowbots" proposed by Bob Kahn and Vint Cerf at CNRI; the deployment of the Z39.50 protocol for searching databases at a level of semantic abstraction by the library community (and beyond); and federated relational databases.

This is now widely used on various kinds of scientific or other scholarly information management and analysis platforms in many disciplines and subdisciplinary communities. The problem is that each of these platforms is a unique, bespoke system specific to the semantics of the data it holds (and the practices of the data community). These platforms are usually expensive to build, maintain and operate.

Increasingly, one sees discussion of two types of data repositories: disciplinary repositories, which often have the characteristics just described, and "generalist" repositories such as Dryad, which essentially host files and metadata describing those files, but where the model is very much one of file transfer out of the repository for use.

In an era of enormous datasets (and also a growing mass of data that is sensitive for one reason or another, and the data holder may allow computations that for example capture statistical characteristics of the data but not want to release the underlying data) it seems clear that it's desirable to find ways to support transfer of computation to data resources more generally; put another way, this would greatly extending the capabilities of generalist repositories.

It's interesting to note that the research data management community has already recognized this need in the abstract through the adoption of the FAIR (Findable, Accessible, Interoperable and Reusable) principles; however, there are great difficulties in understanding how these principles can be turned into effective practice, particularly in resource-constrained environments. A vision of what might be possible here can be found in Barend Mons' recent paper http://www.data-intelligence-journal.org/p/10/6/?parameter=dis .

I'm interested in getting a better understanding of how FAIR practices can facilitate the construction of extended generalist repositories that can receive computations, and reciprocally how the need to support transfer of computation might shape FAIR practices.