OSDC PIRE Workshop 2013 Competition

Pedro D Bello Maldonado
pbell005@fiu.edu
Florida International University

Matthew Greenway
mgreenway@uchicago.edu
Laboratory for Advanced Computing

June 20, 2013

In recent years, scientists around the world have been able to generate huge amounts of data that grow exponentially every year as newer technologies emerge. However, we are still unable to successfully connect various big data sites that belong to different organizations or research areas.

Cross-disciplinary discoveries are brought by inter-disciplinary research and exchange of data/experiments. The OSDC platform is a growing venue for such data. However, one obstacle is the lack of a tool to discover data content and to allow data exchange, independent of the store format. To this end, we want to provide two types of tools: one that deals with categorizing the data content, by using generic semantics about it (what field it comes from, how it was obtained: experiment, benchmark, measurement, simulation); and another which deals with non-functional aspects of the data, e.g. type of database, geo-location.

The main drawback of data exchange is that schemas and metadata are necessary for the unveiling of the content of the data that is being shared. This problem becomes more relevant when the data is not structured and tagged in a useful way when it is collected. Machine learning techniques that compare and learn from different known datasets could solve such problems using stochastic approaches to the solution. Such techniques would rely on interdisciplinary collaboration in order to broaden the spectrum of training samples.

Researchers in different domains work with different tools that best suit their applications. Data collected is usually stored using the databases available to their grasp. However, as inter-disciplinary work grows, tools to translate from one database paradigm to another become more necessary in order to break the bridge between different disciplines. Consider the next situation. Lab A has experimental data stored using MySQL while lab B has experimental data stored using HBase database. Lab A would like to use the data in lab B

using a format that corresponds to the current database paradigm that they already have. Using a translator that matches the information in HBase databases to MySQL databases would help lab A to obtain the data more easily and conveniently. This approach can be taken using a compiler-like solution where an instance of a translation is stored so it doesnt need to be redoing every time a query is made by a different database paradigm.

We propose the UvA Data Service as a solution to these problems. We plan to register the current OSDC metadata database with the Data Service. This provides OSDC users a way to easily interface preexisting databases with the OSDC public datasets. We will position the UvA Data Service as an integral part of the OSDC work flow.