Geophysical Research Abstracts
Vol. 16, EGU2014-12816, 2014
EGU General Assembly 2014

Can EO afford big data – an assessment of the temporal and monetary costs of existing and emerging big data workflows

Oliver Clements and Peter Walker
Plymouth Marine Laboratory, United Kingdom (olcl@pml.ac.uk)

The cost of working with extremely large data sets is an increasingly important issue within the Earth Observation community. From global coverage data at any resolution to small coverage data at extremely high resolution, the community has always produced big data. This will only increase as new sensors are deployed and their data made available. Over time standard workflows have emerged. These have been facilitated by the production and adoption of standard technologies. Groups such as the International Organisation for Standardisation (ISO) and the Open Geospatial Consortium (OGC) have been a driving force in this area for many years. The production of standard protocols and interfaces such as OPeNDAP, Web Coverage Service (WCS), Web Processing Service (WPS) and the newer emerging standards such as Web Coverage Processing Service (WCPS) have helped to galvanise these workflows.

An example of a traditional workflow, assume a researcher wants to assess the temporal trend in chlorophyll concentration. This would involve a discovery phase, an acquisition phase, a processing phase and finally a derived product or analysis phase. Each element of this workflow has an associated temporal and monetary cost. Firstly the researcher would require a high bandwidth connection or the acquisition phase would take too long. Secondly the researcher must have their own expensive equipment for use in the processing phase. Both of these elements cost money and time. This can make the whole process prohibitive to scientists from the developing world or “citizen scientists” that do not have the processing infrastructure necessary.

The use of emerging technologies can help improve both the monetary and time costs associated with these existing workflows. By utilising a WPS that is hosted at the same location as the data a user is able to apply processing to the data without needing their own processing infrastructure. This however limits the user to predefined processes that are made available by the data provider. The emerging OGC WCPS standard combined with big data analytics engines may provide a mechanism to improve this situation. The technology allows users to create their own queries using an SQL like query language and apply them over available large data archive, once again at the data providers end. This not only removes the processing cost whilst still allowing user defined processes it also reduces the bandwidth required, as only the final analysis or derived product needs to be downloaded.

The maturity of the new technologies is now at a stage where their use should be justified by a quantitative assessment rather than simply by the fact that they are new developments. We will present a study of the time and cost requirements for a selection of existing workflows and then show how new/emerging standards and technologies can help to both reduce the cost to the user by shifting processing to the data, and reducing the required bandwidth for analysing large datasets, making analysis of big-data archives possible for a greater and more diverse audience.