Geophysical Research Abstracts
Vol. 13, EGU2011-9280, 2011
EGU General Assembly 2011

MashMyData: a gateway for scientific visualization and intercomparison of secure, distributed data

Alastair Gemmell (1), Jon Blower (1), Phil Kershaw (2), Stephen Pascoe (2), and Ag Stephens (2)

(1) Reading e-Science Centre, University of Reading, United Kingdom, (2) Centre for Environmental Data Archival, Science and Technologies Facilities Council, United Kingdom

Currently, there are many web portals available which focus on visualizing environmental data, but for comparing datasets and carrying out analyses users turn to desktop applications such as Matlab and IDL. These are powerful, but require users to download large datasets and manually get to grips with low-level data formats and metadata conventions. Until this time there have been a number of factors inhibiting the transfer of basic data analysis and comparison from desktop applications to web portals. These include the size of the datasets involved, their diversity of format and location, and security constraints.

MashMyData is a UK NERC (Natural Environment Research Council) funded “Technology Proof of Concept” programme which aims to address these needs by creating a system to allow environmental scientists to compare and combine diverse datasets over the web without the need to understand the low-level technical details of the data’s format or physical location. Users are able to upload their own data and compare with professionally-curated datasets in data centres, respecting data privacy at all levels.

The technical challenges involved in the MashMyData project have much of overlap with a number of important challenges in the wider environmental informatics community. These are key problems and their solutions will be very widely applicable in future. They can be summarised as a) dealing with data diversity b) performing calculations in a way that scales and c) accessing secure data and handling the delegation problem. In the current presentation we discuss the first two challenges and work we have done in the MashMyData project to address them. The third challenge is the subject of a separate submission (Kershaw et al).

Dealing with data diversity. Data resides in different formats in different physical locations, accessed via different web service protocols. Furthermore, users can upload data in a variety of formats. It is very important to avoid writing specific data processing and visualization code for each individual dataset. The various datasets must therefore be exposed to the rest of the system in a consistent fashion. We have developed a Java implementation of the data model defined by the Climate Science Modelling Language (CSML), which applies international standards to describe a very large proportion of environmental science data (http://ndg.nerc.ac.uk/csml/). The key is that CSML uses a small number of “feature types” to model a large number of datasets. (Feature types are based on the data’s spatiotemporal geometry and include grids, vertical profiles, timeseries, trajectories and points.) All visualization and analysis routines then operate upon these feature types, without knowledge of how or where the underlying data are stored.

Data collocation is a key feature that is needed. This is very difficult to achieve if we don’t use standard means for representing geography and time. This is where adoption of open GIS standards really pays off for scientists. By providing an attractive tool that allows comparison of diverse data we further encourage scientists to adopt these standards.

Performing calculations remotely in a way that scales. In order to ensure future scalability, and to avoid large data transfers where possible, MashMyData demonstrates the processing of data on remote compute servers that are close to the data stores. We have employed the OGC Web Processing Service (WPS) as the interface to the remote compute servers. There is much current community interest in the use of WPS for this purpose, although the technology has rarely been employed in the environmental sciences. This work builds upon previous experience within the project team from engagement with the UK Climate Impacts Programme.