Not Your Average Visualisation Project

J.W. Larson[a] and P.R. Briggs[b]

[a] Mathematical Sciences Institute, The Australian National University
[b] CSIRO Marine and Atmospheric Research
Email: jay.larson@anu.edu.au

Abst This presentation combines a tale of software development with the opportunities for exploration that seemingly mundane—but well-designed—software can offer. We have created an open-source, python-based, and massively parallel visualisation software system, which we employed to create the imagery for the Australian Carbon and Water Observatory (ACWO; http://carbonwaterobservatory.csiro.au/). ACWO imagery is generated from 26 data variable fields, displayed at national, state, catchment, and natural resource management scales, comprising a collection of millions of images. A sampling of this imagery is presented to demonstrate its illuminative character. The system is sufficiently generic that it can be deployed to the Australian Water Availability Project (AWAP) dataset. We then point out the file-and-data-wrangling facilities, combined with python's formidable numerical mathematics and scientific computing software ecosystem, provide a solid foundation on which big data analytics applications for AWCO and AWAP can be built with aplomb. We conclude with some representative case studies.

"Big Data Analytics" or "Data Science" is currently a hot topic in the commercial sector and the disciplines of computational science and engineering. It is widely agreed that the main obstacles on the road from raw data to understanding are assembling data, implementing quality control, understanding file-naming conventions and data formats, and getting the data into the analysis engine; processes collectively called "data wrangling," "data munging," or "data janitor work." Data munging can occupy 50-80% of the total effort involved in the data mining process[1]. CSIRO's Earth Observation Informatics Transformational Capability Platform provided us with the support to create comprehensive imagery for ACWO[2,3] and the opportunity to amortize the ACWO data-wrangling effort to enable deep analysis of the ACWO data.

The ACWO field variables are organised as two-dimensional arrays of data on a 0.05° x 0.05° latitude-longitude grid that covers the Australian region's landmass. Data files contain one 2D slice per time sample, and the time sampling period is typically one month, spanning the period 1900-present. Thus, for a single variable at the national scale, there are over 1300 data files, corresponding to 1300 images. The ACWO web site's display system shows a mosaic of "thumbnail" images, each of which may be clicked to reveal a full-size, high-resolution image. Thus,visualisation of a single field at the national scale a total of over 67600 images. Generating each image on the fly is prohibitively expensive computationally, and all of the imagery was generated en masse on the CSIRO's burnet multiprocessor cluster. The visualisation system is built on open-source Python tools, notably numpy, matplotlib, and matplotlib-basemap. The parallelization is achieved using the Message Passing Interface Library via mpi4py. Animations are produced from these images using the open-source ffmpeg package. Additionally, the system supports generic geographic masking/subsetting of data; at present, state, catchment, and natural resource management area regionalisations are supported, with the ability to implement and apply user-defined regionalisations. Happily, the ACWO and AWAP file format and file/directory naming conventions coincide, and the system has been tested successfully with AWAP data. Thus, the software infrastructure required to perform the "data janitor" work for ACWO and AWAP is now available in a single open-source code base.

The visualisation system will provide multiscale imagery later this year. Case studies employing information-theoretic techniques will be presented to demonstrate the system's potential to extract new information regarding climate variability and climate change.

Keyw *Carbon and Water Analyses, Big Data Analytics, Climate Variability, Climate Change.*

---

[1] S. Lohr, "For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights," *New York TImes, 17 August 2014.* http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0

[2] P. Briggs, V. Haverd, J. Larson and P. Canadell, "Australian Carbon Observatory and Delivery System," Final Report, CSIRO Marine and Atmospheric Research (2013).

[3] Australian Carbon Water Observatory Web Site: http://carbonwaterobservatory.csiro.au/.