BlogMyData: A Virtual Research Environment for collaborative visualization of environmental data

J.D. Blower, A. Santokhee: Reading e-Science Centre, Environmental Systems Science Centre, University of Reading

J.G. Frey, A. Milsted: School of Chemistry, University of Southampton

1. The problem

Understanding and predicting the Earth system requires the collaborative effort of scientists from many different disciplines and institutions. The National Centre for Earth Observation (NCEO) and the National Centre for Atmospheric Science Climate Group (NCAS-Climate) are both high-profile interdisciplinary research centres involving numerous universities and institutes around the UK and many international collaborators. Both groups make use of the latest numerical models of the climate and earth system, validated by observations, to simulate the environment and its response to forcings such as an increase in greenhouse gas emissions. Their scientists must work together closely to understand the various aspects of these models and assess their strengths and weaknesses.

At the present time, collaborations take place chiefly through face-to-face meetings, the scholarly literature and informal electronic exchanges of emails and documents. All of these methods suffer from serious deficiencies that hamper effective collaboration. For practical reasons, face-to-face meetings can be held only infrequently. The scholarly literature does not yet adequately link scientific results to the source data and thought processes that yielded them, and additionally suffers from a very slow turnaround time. Informal exchanges of electronic information commonly lose vital context; for example, scientists typically exchange static visualizations of data (as GIFs or PostScript plots for example), but the recipient cannot easily access the data behind the visualization, or customize the visualization in any way. Emails are rarely published or preserved adequately for future use. The recent adoption of "off the shelf" Wikis and basic blogs has addressed some of these issues, but does not usually address specific scientific needs or enable the interactive visualization of data.

2. The solution

A Virtual Research Environment is an attractive solution to the above problems. In the JISC-sponsored BlogMyData project we are creating such a VRE by combining the capabilities of two existing technologies that have already seen wide adoption among scientists:

1. The Godiva2 data visualization system (http://www.reading.ac.uk/godiva2) provides a means for scientists to browse interactively in a "Google Maps-like" fashion through large environmental datasets, including numerical model outputs and high-resolution satellite imagery, using only a web browser. Scientists can produce maps, timeseries and other plot types. This system completely removes the need for the scientist to understand the technical details of how and where the data are stored.
2. The LabBlog is a web-based blogging tool specifically designed for the practising scientist to record, disseminate and evaluate their research. The Blog can also be used as a collaboration tool that allows discussion between colleagues. For open science work the blog uses standard protocols (such as Really Simple Syndication, RSS) to publish its content to the public domain but also contains the necessary access control to keep any private work secure. Although initially designed for the use of laboratory chemists, the LabBlog is being adapted in this project to meet the needs of environmental scientists.

Having logged in to the BlogMyData VRE using OpenID, scientists examine output from the latest cutting-edge climate and ocean models using the Godiva2 interface. Upon finding a feature of interest (perhaps an extreme event, or a suspected problem with the model) the user creates a new blog entry that is linked to the current visualization. The blog entry is automatically tagged with metadata about the feature of interest (e.g. its location in time and space, and the dataset from which it is derived). Colleagues provide input through comments and by linking blog entries together. Through semantic and geospatial tagging, scientists can discover colleagues working on similar scientific problems. The system is augmented by the addition of a geospatially-enabled database, based on the widely-used open-source PostgreSQL database with the PostGIS

extensions. This database will associate blog entries with geographical areas and time periods and allow users to discover discussions that relate to particular areas of interest very efficiently. See Figure 1.

3. Progress and lessons learned so far

The system is being developed under an iterative process, with regular feedback from users in NCAS-Climate and NCEO. We have created an end-to-end prototype of the system, in which users can create blog entries based upon map-based visualizations (i.e. horizontal x-y views of the data). Blog entries are captured in a private blog, which is only visible to a controlled set of users (authenticated using OpenID), thereby maintaining the privacy of the research. Blog entries are syndicated as GeoRSS feeds (GeoRSS is an enhancement to Really Simple Syndication, in which each entry is tagged with geographic information). These feeds can be consumed in standard RSS viewers (such as Microsoft Outlook, Google Reader and Firefox Live Bookmarks), or in "geo-enabled" viewers such as Google Maps (Figure 2). These feeds provide a simple means for scientists to discover research activity in related areas.

We have tested the prototype on some members of the NCAS-Climate group, who are working on the development of the latest high-resolution climate models, including HiGEM (http://www.higem.ac.uk). This has generated some interesting initial feedback, which will guide the next phase of development. Among the most important items of feedback are:

- The privacy controls are regarded as essential: without these, the users would hesitate before posting their most interesting thoughts.
- Content is king: the VRE must display exactly those data that the users are interested in at the current time (sample "test" data from other domains is much less engaging). We have therefore gone to considerable effort to ensure that the data are relevant and presented correctly[1].
- The generation of animations of data gives scientists a great deal of insight into the dynamics of the Earth system. The Godiva2 system is popular for its ability to generate animations of complex numerical model data quickly and easily. As a high priority, the VRE as whole will be amended to allow blogging about animations of data, not just static images. (We did not initially anticipate this being a high priority for users.)

4. Future work

These are the main areas of technological development that will be performed over the remainder of the project (which will finish at the end of September 2010):

- Adding the ability to blog about more types of visualization (not just maps). We anticipate that users will want to blog about timeseries graphs, vertical sections and profiles, as well as the animations that have been specifically requested. Many different types of visualization are possible and development will be prioritized according to user feedback.
- Adding the ability to create customized GeoRSS feeds of blog entries. These feeds are the primary means by which we envisage that users can discover each other (in addition to browsing the BlogMyData website). Custom feeds will be created based upon geographical area, dataset, geophysical variable or time window (e.g. "Give me a feed with all blog entries about sea water density in the North Atlantic"). In addition, feeds of "hot topics" could be created to allow investigators to monitor the latest activity of their research groups.
- Experimentation with presenting information in different ways: for example the blog entries could be exported as KML (http://code.google.com/apis/kml/documentation/), which is richer than GeoRSS and will allow content to be viewed on Google Earth.

Some of the more sophisticated features are only useful once a significant body of blog entries are present in the system. Therefore we are focusing our efforts on gradually building the user base, working with specific users and attempting to answer their current issues, before making large strides in development.

[1] For example, climate models often use a 360-day calendar in order to increase the consistency of monthly, seasonal and yearly statistics. The system had to be modified to handle this unusual calendar system correctly, a task we had not anticipated.

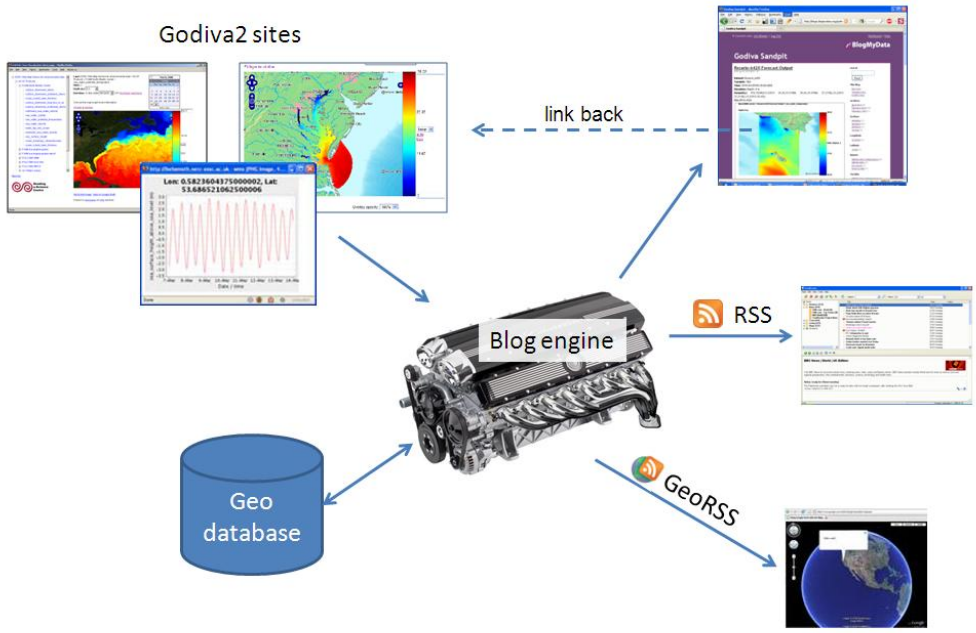

Figure 1: Sketch architecture of the BlogMyData system. Users explore environmental data using Godiva2 sites, which project information onto draggable, zoomable maps. Users create blog entries that are linked to particular visualizations, which are stored in the blog engine, which uses a geospatial database to store geospatial and temporal information. The blog entries are displayed on the project website, on which other users can leave comments. Each blog entry links back to the Godiva2 site that created it, preserving the state of Godiva2 at the time of creation, allowing easy further exploration. Content is syndicated via RSS (for standard feed readers) and GeoRSS (for geo-enabled feed readers).

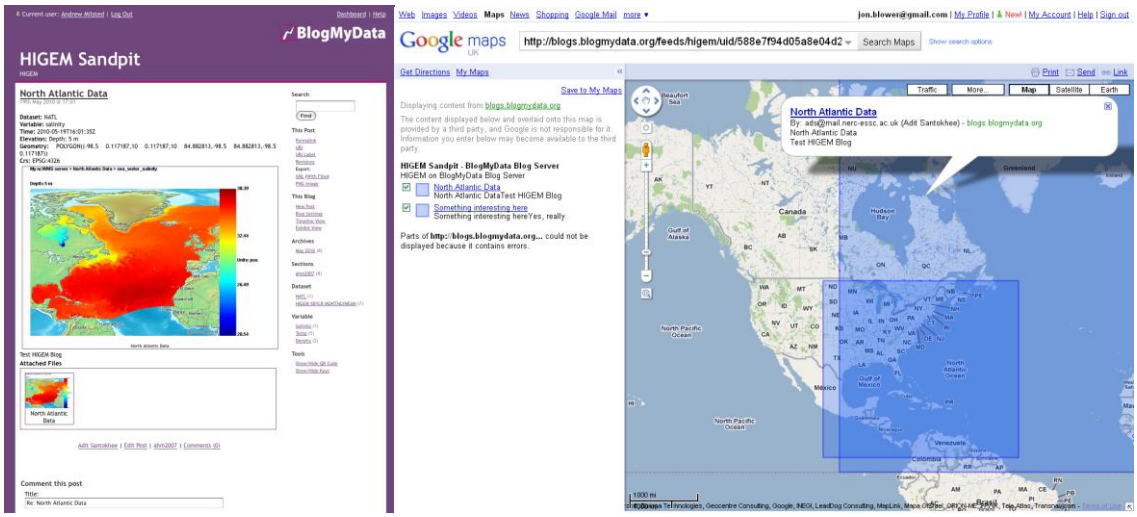

Figure 2: Detail figure showing the display of blog entries on the project website (left) and a GeoRSS-enabled feed reader (Google Maps, right).