A Data Centered Collaboration Portal to Support Global Carbon-Flux Analysis[1]

Deb Agarwal (LBL/UCB), Marty Humphrey (UVa), and Norm Beekwilder (UVa)

Abstract

We have developed a scientific collaboration portal, called *fluxdata*, which now serves the approximately 400 scientists analyzing the global FLUXNET carbon-flux synthesis dataset. The portal is designed to serve three types of users: measurement site scientists, synthesis teams, and the public. Key aspects of our portal design philosophy include: minimize the barrier to entry, focus on functionality targeted at improving the **science** experience, avoid adding extraneous functionality, include critical content needed by a majority of collaborators, listen to the scientists, and anticipate changes in users' practices due to introduction of collaborative tools. The resulting portal provides extensive collaboration and data support through data reports, data download access, a blog, e-mail functions, and notification/update options and has been supporting the FLUXNET users for a little over a year.

FLUXNET

The FLUXNET network is a global network of over 400 carbon-flux measurement sites which provide long-term carbon, water, and energy flux data. The current dataset, which contains 960 site years of data collected by more than 129 scientists working at over 253 measurement sites within FLUXNET, is critical to understanding climate change through cross-site, regional, ecosystem, and global-scale analyses. The FLUXNET measurement sites are managed by individual scientists. Regional networks provide data centers and similar common services. FLUXNET was established as a global set of collaborating regional networks and measurement sites. The regional networks include: CarboeuropeIP, AmeriFlux, Fluxnet-Canada, LBA, Asiaflux, Chinaflux, USCCC, Ozflux, Carboafrica, Koflux, NECC, TCOS-Siberia and Afriflux (1).

Currently there are 70 approved FLUXNET analysis teams (proposals) and this number continues to grow (For the current list see (2)). A typical proposal team involves 2-20 collaborators who together identify the list of measurement sites to use, analyze the data, and write the paper. Before publication of results, the analysis team must contact the scientists running the measurement sites to ask permission to use the data, request any additional information needed, and invite participation in the analysis effort. The FLUXNET synthesis dataset is a living dataset with new additions and updates happening regularly.

Collaboration

We have created a FLUXNET collaboration portal Using SharePoint Server 2007 that supports synthesis activities (http://www.fluxdata.org). There are three primary functions of this portal:

[1] The research and development described in this paper is funded by the Microsoft Corporation. This work is also supported by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

support data access and communication needs of analysis activities, support data update and synthesis information needs of measurement site scientists, and provide public information about the dataset and synthesis activities. Although all these functions involve the data, each requires a unique set of functionality. Our portal design philosophy is based on our team's experience developing collaboration tools for scientific environments (3), (4), (5) and the experiences of other groups developing scientific collaboration environments (6), (7), (8), (9), and (10).

The public section of the portal contains all information about the dataset and collaboration that can be made openly available. It provides information such as characteristics and locations of the measurement sites via interactive maps and reports. It contains lists of the variables measured including the explanations and availability of those variables. A blog contains regular updates, announcements of changes to the dataset, and information about new portal functionality. The synthesis teams' membership, proposals, progress, and lists of the sites involved in each proposal are also included as well as a user manual. Since this area of the portal is publicly accessible, it has the lowest barrier to entry and helps new users get oriented. It also allows potential users to evaluate the expected utility of the dataset before submitting a proposal to use the data.

The remainder of the portal is accessible to authorized users only. Core functions in this part of the portal provide access to the FLUXNET synthesis dataset. This area of the portal is the only place to browse or download the data. Since this data is critical to all synthesis activities, at least one member of each proposal team can be expected to register on the site. The data provides some of 'killer content' that helps drive adoption of the portal.

Each synthesis team is responsible for notifying and working with the measurement site scientists for data they are using. At the regional level, notification of site scientists is typically implemented via an automated e-mail to the site scientist each time the data is downloaded. Since a synthesis effort will generally start from a large number of possible sites (often all 253) and screen out inappropriate sites by downloading and checking the data, this method of notification leads to many false positives and does not scale. In the FLUXNET portal implementation we wanted to significantly improve the information flow between the synthesis teams and the site scientists. Working closely with the users, we developed use cases and defined a new notification model. In our resulting design, each synthesis team is able to specify via the portal the sites used in their analysis effort. A search function is provided that allows a site scientist to learn which synthesis activities are using their site's data.

To reduce spurious downloads of site data, we provide on the fluxdata site available ancillary site information and annual averages for all the variables and sites to help synthesis teams to narrow down their selections before downloading the data. The communication between the synthesis teams and the site scientists is further facilitated via the portal by built in e-mail mechanisms for a proposal team to compose and send e-mail to the measurement sites selected. The portal infrastructure determines the recipients based on the site selections of the proposal team. Synthesis teams often define their site list in order to e-mail sites. Once the site list is defined by a synthesis team, measurement site scientists can see who is using their data. The usage of this feature has been ramping up as the synthesis teams begin serious analysis efforts.

Ideally all the necessary data would be collected before synthesis begins; however, this is not generally possible for ancillary data. Collection of ancillary site information is a long and iterative process. Often ancillary data collection, particularly biological data, is most effectively carried out by synthesis teams trying to write a particular paper. The challenge is to capture and curate this data when it is collected. To support this goal, we have implemented portal functions that allow users to submit corrections and additions to the ancillary data. Data submitted through the portal is verified by a curator before it is accepted. This allows all users to submit ancillary data updates and an expert to curate the submitted data. This model has increased the likelihood that values will get corrected in the system.

CONCLUSION

The FLUXNET collaboration portal has been up and running for over a year and usage of the site has steadily increased over the year. It has become the primary site that users go to for information about FLUXNET sites and data. The site is serving on the order of 400 scientists. This portal provides a prototype and infrastructure that could benefit other data-centric scientific collaborations and a next step will be to apply it to a hydrologic collaboration.

BIBLIOGRAPHY

1. *TURNER REVIEW No. 15, 'Breathing' of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems.* **Baldocchi, Dennis.** 1, 2008, Australian Journal of Botany, Vol. 56, pp. 1-26.

2. Fluxdata approved proposal page. *Fluxnet Synthesis Dataset.* [Online] http://www.fluxdata.org/DataInfo/Dataset%20Doc%20Lib/PaperWritingTeamsInfo.aspx.

3. *The Reality of Collaboratories.* **Agarwal, D. A., Sachs, S. R. and Johnston, W. E.** Berlin, Germany : s.n., April 1997. Proceedings of Computing in High Energy Physics.

4. *Collaboration Tools for the Global Accelerator Network.* **Agarwal, D., Olson, G. and Olson, J.** Berkeley, CA : s.n., August 2002. Final Report of the Collaboration Tools for the Global Accelerator Network Workshop.

5. *A New Security Model for Collaborative Environments.* **Agarwal, D., et al.** Seattle, WA : s.n., June 2003. Proceedings of the Workshop on Advanced Collaborative Environments.

6. *From Laboratories To Collaboratories: A New Organizational Form for Scientific Collaboration.* **Finholt, T. A. and Olson, G. M.** April 2006, Psychological Science, pp. 28-36.

7. *A Science Collaboration Environment for the Network for Earthquake Engineering Simulation.* **Youn, Choonhan, et al.** 2007. Grid Computing Environments Workshop.

8. **Thakar, Ani.** The Sloan Digital Sky Survey: Drinking from the Fire Hose. *Computing in Science & Engineering.* January/February 2008, Vol. 10, 1, pp. 9-12.

9. **Plale, Beth, et al.** CASA and LEAD: Adaptive Cyberinfrastructure for Real-Time Multiscale Weather Forecasting. *Computer.* Nov. 2006, Vol. 39, 11, pp. 56-64.

10. **Kouzes, R. T., Myers, J. D. and Wulf, W. A.** Collaboratories: Doing Science on the Internet . *IEEE Computer.* August 1996, pp. 40-46.