

A DATA-CENTERED COLLABORATION PORTAL TO SUPPORT GLOBAL CARBON-FLUX ANALYSIS¹

DEB AGARWAL (LBL/UCB), MARTY HUMPHREY (UVA), NORM BEEKWILDER (UVA), KEITH JACKSON (LBL/UCB), MONTE GOODE (UCB/LBL), AND CATHARINE VAN INGEN (MICROSOFT)

Abstract

Carbon-climate, like other environmental sciences, has been changing. Large-scale synthesis studies are becoming more common. These synthesis studies are often conducted by science teams that are geographically distributed and on datasets that are global in scale. A broad array of collaboration and data analytics tools are now available that could support these science teams. However, building tools that scientists actually use is hard. Also, moving scientists from an informal collaboration structure to one mediated by technology often exposes inconsistencies in the understanding of the rules of engagement between collaborators. We have developed a scientific collaboration portal, called *fluxdata.org*, which serves the community of scientists providing and analyzing the global FLUXNET carbon-flux synthesis dataset. Key things we learned or re-learned during our portal development include: minimize the barrier to entry, provide features on a just-in-time basis, development of requirements is an on-going process, provide incentives to change leaders and leverage the opportunity they represent, automate as much as possible, and you can only learn how to make it better if people depend on it enough to give you feedback. In addition, we also learned that splitting the portal roles between scientists and computer scientists improved user adoption and trust. The fluxdata.org portal has now been in operation for ~1.5 years and has become central to the FLUXNET synthesis efforts.

1. BACKGROUND

A global network of over 400 carbon-flux measurement sites provides long-term carbon, water, and energy flux measurement data for a broad array of ecosystems. Each measurement site is operated by a team of scientists studying the local carbon flux and micro-meteorological properties along with other related local processes. These measurement scientists typically coordinate data collection with and send their data to a regional network for archiving and distribution to scientific analysis efforts. The regional networks include: [CarboeuropeIP](#), [AmeriFlux](#), [Fluxnet-Canada](#), [LBA](#), [Asiaflux](#), [Chinaflux](#), [USCCC](#), [Ozflux](#), [Carboafrika](#), [Koflux](#), [NECC](#), [TCOS-Siberia](#) and [Afriflux](#) (1).

Carbon-climate measurement site data encompasses sensor measurements, field observations, laboratory analysis of field samples, as well as anecdotal descriptions. Each FLUXNET measurement site deploys a number of sensors measuring micro-meteorology such as precipitation or wind speed and carbon flux. The field data are processed by the

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

measurement team to generate ½ hourly *flux-met* data files. The other, *ancillary data* include relatively infrequent measurements of variables such as soil carbon, leaf out dates, vegetation species, canopy height, soil characteristics, and site disturbances such as planting or wildfire. An important aspect of the ancillary data is that each measurement is often annotated. For more information about ancillary data and related reporting protocol see (2).

FLUXNET is a collaboration between the regional networks and independent measurement sites which results in a global carbon-flux measurement network. This global network is brought together to enable global-scale synthesis activities. Each FLUXNET dataset is gathered in preparation for a global synthesis workshop. The aim of each workshop is to produce a new FLUXNET dataset, discuss its scientific potential, propose new cross-site synthesis analyses and start these analyses. The first FLUXNET workshop was held at the Marconi Conference Center in 2000. Before the Marconi workshop a dataset was gathered that contained 97 site years of data from ~40 sites that were located primarily in the Americas and Europe. The dataset was quality checked and gap filled. The synthesis analysis efforts using this Marconi dataset resulted in 11 synthesis papers published in a special issue of *Agricultural Forest and Meteorology* in December 2002.

In the case of the Marconi dataset, all the synthesis teams knew the measurement site scientists personally so communication and trust between the two groups was relatively easy to establish. The data download and communication functions were handled manually for the Marconi dataset. In addition, the dataset contained only the flux-met data from each site. The synthesis teams were responsible for working with the measurement site teams directly to obtain any needed ancillary data for the site.

The second FLUXNET synthesis workshop was held in La Thuile, Italy in February, 2007. The La Thuile workshop was attended by 60 scientists, who represented a cross section of FLUXNET measurement sites. Although an attempt was made to gather and process (quality check and gap fill) the dataset in advance of the workshop, the data was not ready at the time of the workshop. As a result of the workshop and continuing efforts since the workshop, the La Thuile dataset was released to the synthesis teams in September 2007. The delay in releasing the data was due to both the time needed to complete the gathering and processing of the data and the time needed to develop a portal to support sharing of the dataset. The released La Thuile dataset currently contains over 960 site years of data from more than 180 scientists working at over 250 measurement sites around the world. Figure 1 shows the tower locations of the sites contributing to the La Thuile dataset.

The LaThuile dataset is an order of magnitude bigger than any carbon flux measurement dataset that has been available before and will enable cross-site, regional, ecosystem, and global-scale analyses. When compared to the Marconi dataset, this dataset is an expansion by roughly an order of magnitude in number of site years (967 vs 97), number of sites (253 vs 40) and number of measurement scientists involved (180 vs ~30), The LaThuile dataset is also richer than the Marconi dataset because the ancillary data is considered part of the dataset. The number of dataset users is also larger (73 vs 11 paper proposals involving more than 136 scientists) and these numbers all continue to grow (3). As a result of this growth, only a few measurement site scientists if any know all the synthesis teams.

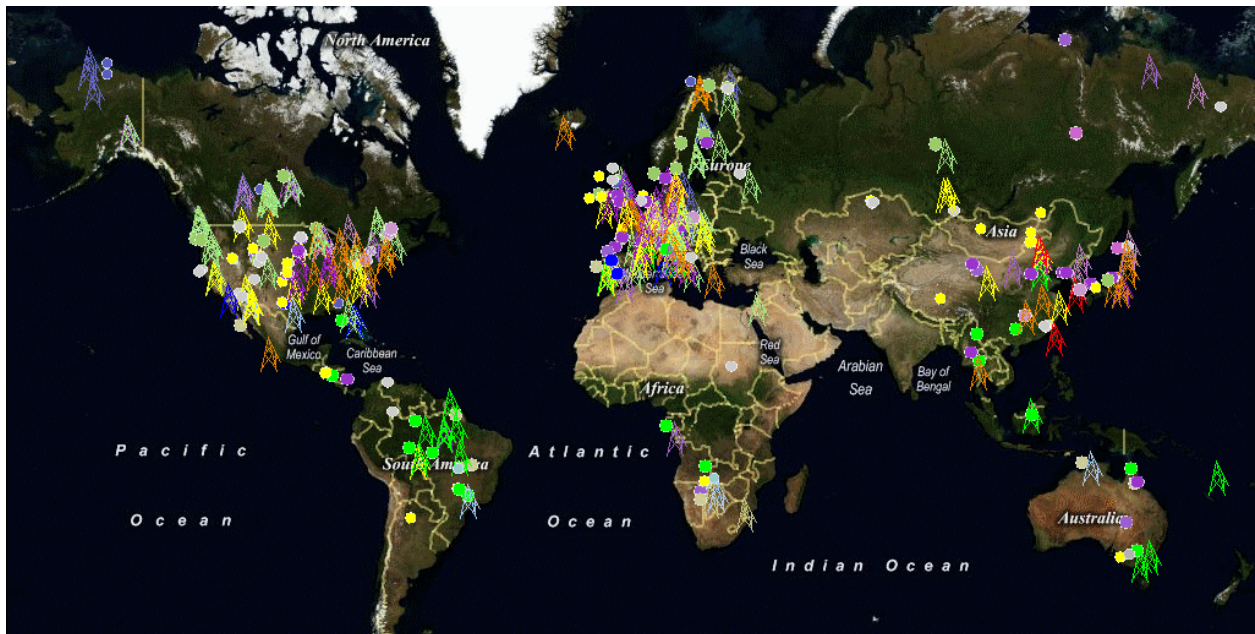


FIGURE 1: MAP SHOWING THE FLUXNET TOWERS CONTRIBUTING TO THE LA THUILE FLUXNET SYNTHESIS DATASET (MARKED WITH THE TOWER ICONS).

The use of the La Thuile dataset is governed by a data fair use policy and is not intended for open public access. A steering committee that includes representatives of the measurement site scientists, regional networks, and synthesis teams was formed to manage the FLUXNET La Thuile synthesis activities. Each team that wanted to use the data submitted a proposal to the steering committee. A typical synthesis team involves 2-20 collaborators who originate a synthesis proposal, identify the list of measurement sites to use, analyze the data, and write the paper. The data fair use policy defines expected interactions between the synthesis activities and the measurement scientists. In particular, it requires that before publication of results, the analysis team must contact the scientists at the measurement sites used in the analysis and inform them of the data usage, confirm permission to use the data, request additional information needed for the analysis, invite participation in the analysis effort, and obtain proper citations and acknowledgments for the data. The La Thuile data fair use policy (4) and the steering committee are designed to protect the interests of the measurement site scientists.

Data for a measurement site scientist represents a potential revenue stream in that it enables analyses that can be carried out using the data (5). For the data contributors to the La Thuile dataset, the potential reward is that they get co-authorship or citation of their data but the risk is that the data contribution will not be acknowledged or will be misinterpreted. The measurement site scientists' conditions for sharing their data with synthesis activities are that: they receive proper "credit" for their data contributions, they be given an opportunity to explain peculiarities of their data, and their own local analysis efforts not be "scooped" by the synthesis activity. Since the synthesis teams are all expected to include at least one person affiliated with a measurement site, there is an assumption that the synthesis teams understand the importance of giving the data providers credit.

Early in the process of gathering the dataset for La Thuile it became clear that new more automated data analysis and synthesis support infrastructure would be required to support the La Thuile synthesis activities. The scale of La Thuile and the need to enable the building of trust between participants meant it was no longer possible to rely on informal and manual mechanisms to manage the communication between synthesis teams and the measurement site teams. The analysis process itself also needed to be revamped for the La Thuile dataset since few of the synthesis teams had the programming skills required to automate the analysis of the data files. In addition, the Marconi dataset practice of having the ancillary data collected by each individual synthesis effort was widely recognized as inefficient and problematic. Measurement site scientists were being asked for the same ancillary data many times and often the ancillary data held by different synthesis teams did not match. Methods for centrally collecting, storing, and presenting the ancillary site data were needed. *The FLUXNET community was desperate for new synthesis support tools for the La Thuile dataset.*

2. THE FLUXDATA.ORG COLLABORATION PORTAL

The first step in designing a portal for FLUXNET was to enumerate and understand the requirements of the expected users of the portal. We had already been working with the AmeriFlux regional carbon flux network on developing a data analysis infrastructure for use in analyzing their dataset (6). That infrastructure provides advanced data organization, mining, and analysis features through utilization of a database to organize the data and On-Line Analytical Processing (OLAP) data cubes for browsing the data (7). Our work with AmeriFlux infrastructure provided a basis for the work with the groups involved in the La Thuile dataset.

2.1. PORTAL REQUIREMENTS GATHERING PROCESS

As a first step in the portal design process, we worked with the La Thuile scientists to develop use cases to allow us to better understand the necessary functionality. There are four primary types of users and associated use cases that must be served by the infrastructure. These are

- Synthesis scientists (data users) – site selection, dataset information data download, analysis support, and paper writing support
- Measurement site scientists (data providers) – proposal information, data download, and data update
- Regional flux networks (data curators) – data correction, checking, and update
- The public – proposal information, dataset information

Although it is tempting to think of each of the above user groups as being distinct, this is not the case. Many use cases are shared by multiple groups. Also, many of the measurement site scientists are involved in synthesis activities and regional networks.

A brief description of each of the primary use cases we developed is provided below:

- Synthesis site selection – evaluate criteria that will determine which sites are suited to an analysis. Typically most of the site selection process is done using high-level aggregated data about the sites.

- Dataset information – ability to quickly answer simple questions about the dataset such as which sites are included, which years of what data, where are the sites located, and who is the measurement scientist in charge of a site.
- Data download– ability to browse and download flux-met and ancillary data for sites. Provide only authenticated access to data.
- Data update – submit updates to ancillary data and new ancillary data and track provenance of ancillary data.
- Analysis support – browsing of aggregated values for sites and compilations of ancillary data and interact with data using MATLAB and Excel
- Paper writing support – enable communication with measurement site scientists and gathering of citations and acknowledgements for data.
- Proposal information – access to proposal theme, progress, and site selection.
- Data curation – inform curator of submitted changes and provide an opportunity for a person familiar with the site, the *curator* to sanity check data submissions.

The fluxdata.org collaboration portal (<http://www.fluxdata.org>) was developed to support the above users and use cases. As part of the design process we attended the science meetings of the groups involved and discussing the use cases with a broad array of the scientists involved. The fluxdata.org design team also participated in the FLUXNET La Thuile meeting and read a representative set of the existing carbon flux papers to better understand the terms, issues, and types of analyses undertaken. These interactions within the community also helped us to build relationships with the scientists and get to know the community. This approach allowed us to understand the current state of practice for the community and to work with them to define scalable, technology mediated practices that could serve as a starting point for a portal that could easily be adopted by the community. As the synthesis activities and the system have matured, the requirements and use cases have continued to expand and we add features to the portal as needed to keep pace.

The next several sections describe the resulting fluxdata.org collaboration portal developed to support the La Thuile synthesis users and use cases.

2.2. *THE PUBLIC ACCESS AREA OF THE PORTAL*

The public area of the fluxdata.org portal can be reached by all users without any restrictions or identification. The public area contains all information about the dataset and collaboration that can be made openly available. This public information is designed to be access by all users so its content is not replicated in other areas of the portal. The public area of the portal contains:

- Characteristics and locations of the measurement sites along with information about the science teams running the measurement sites. This information is presented using interactive maps and reports. Examples include an interactive mashup of tower locations, and reports containing the average annual values of the micro-meteorological flux-met data for the site.
- Lists of the variables measured at sites including the explanations and availability of those variables along with explanations of the derived variables, gap filling techniques, aggregation method, and quality markers.

- Measurement site pages listing all public information about the site and data from the site including pictures of the site if available.
- A blog providing 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 synthesis analysis.
- A user manual describing how to use each of the site's functions.
- Data fair use and publication guidelines.

This area of the portal helps new users get oriented. It also allows potential users to use the publicly available information about the data to evaluate the expected utility of the dataset before submitting a proposal to use the data. This enables a low barrier to entry for new users. This part of the portal has been in operation since late 2007 and in the period January 1, 2008 – March 31, 2009 it averaged ~11,000 visits/day. In the first three months of 2009 it has averaged ~15,000 hits/day (search engines and our own traffic to maintain the site have already been factored out of these statistics). These visits correspond to an average of 44-50 unique users each day. Figure 3 shows the daily visits for the period.

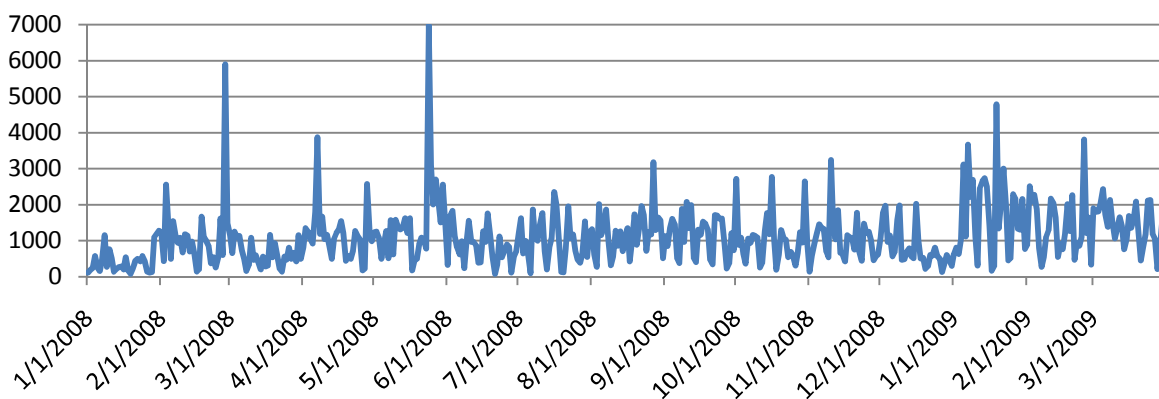


FIGURE 3: FLUXDATA.ORG PUBLIC AREA WEB HITS

2.3. THE SYNTHESIS TEAM SUPPORT AREA OF THE PORTAL

The area of the portal that supports the synthesis teams is accessible only to users that are a member of a proposal and have an account. It provides users with extensive access to the complete La Thuile dataset. Functions available in this area of the portal include:

- Download of flux-met data.
- Browsing of data via MATLAB and Excel.
- Browsing and download of compilations of ancillary data for a site and across sites.
- Access for proposal teams to update the status of their proposal (allows measurement site scientists to track progress), update the list of measurement sites used in the analysis, and exchange e-mail with measurement site scientists.
- Access reports containing annual aggregates of all site flux-met data as well as cross-site compilations of that same data and data quality indicators.

The functions in this area of the portal were the most difficult to design. The users had little experience working at the scale of the La Thuile dataset and sometimes asked for features that would not scale. One example of this is the site identification and communication features. The state of the art in the regional carbon flux networks was for an automated e-mail to be sent to the contributor of the data each time a user downloaded data. This requirement, in the case of the La Thuile dataset, would have led to a large number of notices of downloads that did not result in data usage. Most synthesis efforts started by downloading all the data (over 960 site years) and narrowed the sites down during the analysis process. After working through use cases with the measurement site scientists, we were able to determine that the capability the measurement scientist most wanted was an ability to find out what teams were using their data. In fluxdata.org, each synthesis team can specify sites whose data are used in an analysis. Our design replaces download notification with an opportunity to input information about data usage. It allows measurement scientists to see what analyses have indicated usage of their data. This is just one example where use cases and interaction with the scientists was required to determine the requirement.

In addition, we provided features in the portal such as reports containing compilations of ancillary site information and annual aggregates for the variables to enable synthesis teams to narrow down their site selections without having to download and analyze the data itself. A synthesis effort will often screen out sites based on this ancillary and aggregated site data such as predominant plant species, soil texture, leaf area index, climate, annual rainfall, etc. We have even received anecdotal evidence from synthesis teams of science results being generated off this information.

We realized early in the project that it would be difficult to get synthesis teams to specify their site selections so we provide an incentive. If the synthesis teams specify the sites used in their analysis they can then use an e-mail tool to compose and send e-mail to the scientists running the measurement sites selected by their proposal. The portal automatically determines the e-mail recipients (typically 40-120 scientists). Some synthesis teams have only defined their site list because they wanted to e-mail sites. The usage of this feature has been ramping up as the synthesis teams begin serious analysis efforts. Besides site selection, the synthesis teams have an option to identify the stage their analysis effort is in. Proposal teams were recently asked to use this feature to communicate their status to the steering committee and many of the proposals complied. As of April 2009, there are 73 approved proposals: all but 5 of the proposal coordinators have accounts at fluxdata.org, 42 of the proposals have specified their status, 23 have indicated their site selection, and 47 e-mails have been sent using the e-mail feature. The usage of these features has been slowly increasing over time but is still lower than expected.

2.4. THE MEASUREMENT SITE SCIENTIST SUPPORT FUNCTIONS

The area of the portal that supports measurement site scientists is accessible only to users that are listed as a contact for a measurement site and have an account on fluxdata.org. Site scientists have extensive access to the data from their site. Functions available in this area of the portal include:

- Download of the flux-met data for the site.
- Display all ancillary data collected for the site.
- Submit new ancillary data and update existing ancillary data.

- Search which synthesis activities have specified they are using a site's data.

The original assumption was that measurement site scientists would need no access to the fluxdata.org site and that the ancillary data would be collected along with the flux-met data. But, critical site ancillary data is still being gathered two years after the workshop and we received over 13,000 new or updated ancillary data values between December 2007 and March 2009. One issue with collecting ancillary data was that there were few protocols defined for submission of this data, at the time of the La Thuile workshop. Since the workshop, the reporting protocols have been developed (2) but the diversity of this data still makes consistency difficult. Reporting of data ranges, approximate values, and qualifiers in lieu of simple numeric values is a common practice so we store all ancillary data as text fields to preserve accuracy information and support non-numeric entries.

The submission date, user, and method are captured when ancillary data is submitted. This provenance information is stored along with the value. In addition, all past values and their provenance information are kept so that the history of the values provided for a variable can be reconstructed. This enables ancillary data views that correspond to the database state at any particular point in time to be constructed.

We worked with the scientists to design web interfaces and protocols for reporting values; however, the progress of ancillary data reporting was continually well below expectations. The opportunity to have a site included in an analysis paper turns out to be excellent motivation for the site to submit additional ancillary data. The challenge is to capture that ancillary data in the portal as it is collected by the analysis team. We had hoped that allowing any authenticated user to submit updates and additions to the ancillary data would help but it had little impact. Recently one of the synthesis teams requested that all the sites in their analysis submit specific ancillary data to fluxdata.org, the result was a deluge of ancillary data (1300 ancillary data values in about 2 weeks). This data had been requested many times in general calls to the sites but was not provided until this synthesis paper requested it. This synthesis team acted as a "change leader" based on two incentives we provided: the ability to easily send e-mails to sites and the fact that the portal would automatically get the entries curated and that we were then able to produce customized reports containing the resulting ancillary data collection.

2.5. DATA CURATION, VERSIONING AND RELEASE

Ancillary data submitted to fluxdata.org is stored in a submit table until it has been verified by a curator for the site. The curators have a web interface that allows them to check each submitted value and accept it into the database or reject it. This allows all users to provide ancillary data updates and an expert to curate the submitted data. This model was designed to increase the likelihood that values will get corrected in the system since the person who noticed the problem or got the data is empowered to input the correction or the new data and the data is verified before it is accepted. The primary usage of this feature to date has been by the measurement site scientists and occasionally by a person who has collected a variable (e.g. UTC offset or climate) for all sites.

Data releases are another critical element of the use cases. The La Thuile dataset is often receiving new and updated data. The flux-met data arrives in large ~yearly batches and the ancillary data is more continuously updated. The synthesis teams need to be able to indicate a version of the data that they used in performing their analysis. There are a wide

number of models we can draw from in designing a data release strategy. Agencies such as USGS run their data through processing once from collection to quality checking and release. Released data does not change. NASA uses a collection abstraction for their releases. They continually add new data to a collection as it is received and processed. They reprocess the entire dataset if they revise the processing calibrations or algorithms. Reprocessed data is released as a new collection. For the fluxdata.org site we have adopted a strategy similar to the NASA collections and release a frozen version of the data before each major update to the flux-met data.

3. IMPLEMENTATION OF THE PORTAL INFRASTRUCTURE

We chose Microsoft SQL Server 2005 as the back-end (centralized) data repository and Microsoft Office Sharepoint Server (MOSS) 2007 as the Web server platform. We chose SQL Server 2005 based on its long track record (although not for “scientific” data), scalability, OLAP capabilities, and our personal experience with the platform. We chose MOSS 2007 for its track record in a business context (notably, we had no prior experience with MOSS 2007 before we started this project). To our knowledge, this is the first time that MOSS 2007 has been used to meet the requirements of a scientific collaboration centered on data. We target Internet Explorer and Firefox browsers to ensure wide applicability and interoperability for the users.

We use a fully normalized database schema; each data value occupies a unique data table row. This enables us to add new variables when needed and to build OLAP data cubes directly from the database. The disadvantage of the fully normalized schema is that queries are less intuitive but, we export tabular views to simplify the code needed in the MOSS portal. We use data cubes to generate spatial and temporal aggregations that feed into the summary reports and data products. Simple (e.g. sum, min, and max) and complex (e.g. variance, formulas, or units conversions) aggregations can be computed. Hierarchies for simple filtering provide drilldown capability into each dimension. Client tool integration is evolving, Excel PivotTables allow simple data viewing and we have enabled more powerful analysis and plotting using MATLAB.

MOSS 2007, the Web Server platform, is layered upon Windows Sharepoint Services (WSS), which is itself layered on Microsoft Internet Information Services (IIS). WSS and IIS provide the basic Web portal capabilities, including the ability to create multiple web sites with different security models. MOSS 2007 adds search, basic collaboration, “business intelligence”, and “enterprise content management”. MOSS 2007 is a sophisticated platform with a wide range of functionalities that readily mapped to the requirements that we established (and described in Section 2). Active Directory (AD) provides our account manager and authentication source for the portal. Each user has a unique log-in and group membership(s) (e.g., site scientist, curator, synthesis team, and administrator). We leverage MOSS 2007’s built-in ability to customize the server content based on AD identity and/or group to provide functionality based on role (e.g., only curators see the “curator functionality”). Because each authenticated user automatically has a “Web space” in MOSS 2007, each owner can store metadata regarding his/her data specifically on “their space”, which is then searchable by the MOSS 2007 built-in search server. We have also built in MOSS 2007 a small number of specific functionalities (called “Web Parts”) to support the requirements. These Web parts are:

1. Download my data

2. Submit updates about my data
3. Submit changes to ancillary data
4. Review/Approve/Disapprove submitted changes to data (curate)
5. Surface data releases and accompanying documentation
6. Make an account request
7. Inform authors of data use and/or ask questions regarding the data
8. Invite data authors to participate in scientific exploration/experiment
9. Download data for scientific exploration/experiment
10. Visually inspect data cube(s) via browser

Through the combination of MOSS 2007 and custom web parts, we are able to provide an interface that supports the broad range of services needed by the collaboration and is easily maintainable.

4. RELATED WORK

Although many collaboration portals have been designed to support science, many of them are not adopted by the intended users or are quite limited in functionality. There have been many attempts to understand the underlying drivers and practices that lead to a successful collaboration. The best known of these was a five year study led by the University of Michigan and Howard University into the Science of Collaboratories (8). This study and others conducted by the University of Michigan group resulted in the development of a taxonomy of collaboratories. According to this taxonomy, the FLUXNET collaboration is somewhere between a **Community Data System** (An information resource that is semi-public, of wide interest, and created, maintained, or improved by a geographically-distributed community) and a **Distributed Research Center** (An attempt to aggregate scientific talent, effort, and resources beyond the level of individual researchers) (9).

Scientific data are a source of “monopoly rents” for the owner of the data where revenue accrues in the form of publications, reputation, grants, and students (10). Potential revenues from the data are derived from:

1. Analyses the data owner plans to do on their own.
2. Analyses the data owner plans to do in collaboration with others.
3. Analyses that another scientist could perform that use the data in ways that are uninteresting to the data owner.
4. Analyses that the data owner is interested in but did not think of themselves.

Cases 2-4 have the potential risk that the data owner does not in fact obtain “revenue”. Also, in each analysis there is the possibility that the result might reveal something embarrassing to the data owner. These act as disincentives to sharing of data and must be addressed in a successful collaborative data sharing environment (5).

Our portal design philosophy is based on our team’s experience developing collaboration tools for scientific environments (11), (12), (13) and the experiences of other groups developing scientific collaboration environments (14), (15), (16), (17), and (18).

Many freely available collaboration support environments are now widely available and in use to support scientific collaborations including HUBzero (<http://hubzero.org/>), twiki (<http://twiki.org/>), and droopal (<http://drupal.org/>). These tools are all designed to

provide a shared web environment and have been used successfully by scientific collaborations. However, they would have required extensive development before to add support for features essential to FLUXNET such as dynamic authorization of users, support for the data handling, and support for Excel among others (19) (20).

5. CONCLUSION

The FLUXNET collaboration portal has been up and running for over a year and usage of the site has steadily increased over that time. 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 and to expand the FLUXNET support.

In early 2009 a third FLUXNET synthesis workshop was held in Asilomar. The purpose of the Asilomar synthesis efforts will be to open up the dataset to modelers and enable their use of the data. This new Asilomar dataset will be gathered and released to the approved synthesis groups before the La Thuile synthesis efforts are complete. This overlap in the two efforts will provide us an opportunity to incorporate features that will also support the La Thuile publication effort. A critical part of adding support for the Asilomar dataset to fluxdata.org will be including more publication support features. The next major function planned for the public portion of the portal is a cross referenced paper publication tracking section that will track the papers produced by the synthesis effort and the site years that contributed to the paper's analysis.

6. 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. *Terrestrial Carbon Observing Protocols for Vegetation Sampling and Data Submission.* **Law, B. E., Arkebauer, T., Campbell, J. L., Chen, J., Sun, O., Schwartz, M., van Ingen, C., Verma, S.** 2008. <http://www.fao.org/gtos/doc/pub55.pdf>.
3. Fluxdata approved proposal page. *Fluxnet Synthesis Dataset.* [Online] <http://www.fluxdata.org/DataInfo/Dataset%20Doc%20Lib/PaperWritingTeamsInfo.aspx>
4. FLUXNET La Thuile Data Usage Terms of Reference. [Online] http://www.fluxdata.org/DataInfo/Dataset%20Doc%20Lib/FLUXNETsynthesis_ToR.pdf.
5. *Data at work: Supporting sharing in science and engineering.* **Birnholtz, J and Bietz, M.** 2003. Group.
6. *A Next Generation Flux Network Data Server.* **Agarwal, D, Baldocchi, D and van Ingen, C.** Vienna : European Geophysical Union General Assembly, 2007.
7. *Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals.* **Gray, J, et al.** (1995-11-18). Proceedings of the 12th International Conference on Data Engineering. IEEE. pp. 152-159.
8. *The Science of Collaboratories Project.* [Online] [Cited: March 30, 2009.] <http://www.scienceofcollaboratories.org/>.

9. *From shared databases to communities of practice: A taxonomy of laboratories.* **Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E., et al.** 2, : s.n., 2007, Journal of Computer-Mediated Communication, Vol. 12.
<http://jcmc.indiana.edu/vol12/issue2/bos.html>. 16.
10. **Whitley, R.** *The Intellectual and Social Organization of the Sciences (Second Edition)*. Oxford : Oxford University Press, 2000.
11. *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.
12. *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.
13. *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.
14. *From Laboratories To Collaboratories: A New Organizational Form for Scientific Collaboration.* **Finholt, T. A. and Olson, G. M.** April 2006, Psychological Sci., pp. 28-36.
15. *A Science Collaboration Environment for the Network for Earthquake Engineering Simulation.* **Youn, Choonhan, et al.** 2007. Grid Computing Environments Workshop.
16. **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.
17. **Plale, Beth, et al.** CASA and LEAD: Adaptive Cyberinfrastructure for Real-Time Multiscale Weather Forecasting. *Computer*. Nov. 2006, Vol. 39, 11, pp. 56-64.
18. **Kouzes, R. T., Myers, J. D. and Wulf, W. A.** Collaboratories: Doing Science on the Internet . *IEEE Computer*. August 1996, pp. 40-46.
19. **Humphrey, M., Agarwal, D. and van Ingen, C.** *Publication and Curation of Large-Scale Shared Environmental Scientific Data*. s.l. : Microsoft Technical Report, 2008. MSR-TR-2008-93.
20. *The protection of information in computer systems.* **Saltzer, J. H. and Schroeder, M. D.** 1975. Proceedings of the IEEE. Vol. 63 (9), pp. 1278-1308.