Preliminary version. Final version appears In Proceedings of the 5th IEEE international Conference on e-Science (e-science 2009). Dec 7-9, 2009. Oxford, UK

Fluxdata.org: Publication and Curation of Shared Scientific Climate and Earth Sciences Data

Marty Humphrey*, Deb Agarwal**, and Catharine van Ingen***

*Department of Computer Science, University of Virginia, Charlottesville, VA USA
**Lawrence Berkeley National Lab and Berkeley Water Center, Berkeley, CA USA
***Microsoft Research, Microsoft Bay Area Research Center, San Francisco, CA USA

***Abstract*—Many of today's large-scale scientific projects attempt to collect data from a diverse set of sources. The traditional campaign-style approach to "synthesis" efforts gathers data through a single concentrated effort, and the data contributors know in advance exactly who will use their data and why. At even moderate scales, the cost and time required to find, gather, collate, normalize, and customize data in order to build a synthesis dataset can quickly outweigh the value of the resulting dataset. By explicitly identifying and addressing the different requirements for each data role (author, publisher, curator, and consumer), our data management architecture for large-scale shared scientific data enables the creation of such synthesis datasets that continue to grow and evolve with new data, data annotations, participants, and use rules. We show the effectiveness of our approach in the context of the FLUXNET Synthesis Dataset, one of the largest ongoing biogeophysical experiments. Keywords-escience; collaboration; Sharepoint**

I. INTRODUCTION

The era of remote sensing, cheap ground-based sensors and Web-based access to agency repositories, such as are provided by USGS, NOAA, and NASA, is here. Large-scale virtual organizations such as CASA/LEAD [1], National Virtual Observatory (NVO) [2], CUAHSI HIS [3], and BIRN [4] give scientists new and easier ways to access data and collaborate over the Internet. Many of these large-scale scientific projects are *synthesis* efforts, which by definition attempt to bring together and utilize data from a diverse set of sources and disciplines. In most cases today, these synthesis datasets are gathered in a "campaign" fashion, whereby data is gathered through a single concentrated effort specifically for the synthesis project and the data contributors know in advance exactly who will use their data, what purpose their data will be used for, and the agreed usage rules. When there are a relatively small number of data sources, and these data sources are fairly static and generally similar in quality and form when compared to each other, the campaign style can be an effective process for scientific explorations.

However, large-scale scientific efforts are increasing aimed at addressing much larger problems of much more significance and importance. The campaign-style of focusing on a single fixed point in time will no longer be effective, as problems such as global warming can require the collection of data from a large, global and diverse set of data contributors. Although extensive measurement data has been collected, it is often not directly useful in its native form. With today's mechanisms, even at moderate scales, the cost and time required to find, gather, collate, and customize data in order to build a synthesis dataset can easily outweigh the value of the resulting synthesis dataset – e.g., upon completion of the new synthesis dataset, new data will have become available, from new contributors, with new consumers, with new and diverse usage rules, etc. Simply, for many scientific activities, it is no longer possible or desirable to collect and publish these datasets in this way.

In this paper we describe a novel data publication and curation infrastructure to support management of large-scale shared scientific data such as synthesis datasets. It enables scientific data to be collected, used, and maintained in a sustainable manner and is applicable across a broad array of science domains. The dataset can grow and evolve over time, incorporating new data, new annotations on the data, new participants, and even new use rules. A key element of the design requirements for this infrastructure is to explicitly identify and address the different requirements for each data role (based roughly on the terminology of [5]): The *author* is the producer and contributor of the raw data and wants the data to be used to advance science (but with proper attribution). A *curator* is tasked with maintaining the integrity of a collection of authors' data within the virtual organization. The *publisher* creates the central starting point (e.g., Web portal) to obtain/search the virtual organization's data. The *consumer* is the scientist pursuing an investigation that needs the data. We explicitly separate the concerns/requirements for each role in collecting, publishing, and using a dataset and create a set of generic Web-based software capabilities for each role.

In the Grid community, while there has been some important attention paid to meta-data creation and management (e.g., see [6] for a survey of data provenance systems), most efforts have focused on the issues involved with a scientist attempting to find relevant distributed data for a particular experiment about to be performed [7][8], perhaps through a file system abstraction[9][10]. A fundamental problem in the Grid community in general is that it is generally assumed that the providers of "raw" data are as technically savvy (or even the same people) as the computational scientists who form the hypotheses to explain that data and/or perform sophisticated computational experiments/simulations based on the raw data. Typically, the provider of the raw data (e.g., sensor data) has expertise in the mechanical devices to measure (environmental) phenomena and in the particular domain

science being studied but rarely has the interest, expertise, or time necessary to learn new software designed implicitly only to meet the requirements of computational scientists. The challenge our data management architecture addresses is the ability to provide explicit support for each of the roles, thereby facilitating the data to be collected, processed, and utilized in an evolvable manner while respecting and protecting the roles of the individuals involved.

We demonstrate the utility of our architecture in the context of the FLUXNET [11] Synthesis Dataset, which is one of the largest ongoing biogeophysical experiments based on the number of sites and the site years of data (~140 data contributors and ~80 researchers using the data set). We developed and maintain the infrastructure and serve as the top level data curators and publishers for the global FLUXNET synthesis dataset which currently uses the infrastructure to host the dataset.

The remainder of this paper is as follows. In Section 2, we establish the requirements, both on a per-role basis and for the system as a whole. In Section 3, we describe our system architecture, the initial data representation (Section 3.1) and our Web-based support for each of the roles (Section 3.2). In Section 4, we evaluate our approach and describe how we apply these principles to the FLUXNET Synthesis Dataset. Section 5 concludes.

II. REQUIREMENTS

We established the goals and requirements of the data architecture by considering the desired properties of the system as a whole as well as the requirements on a per-role basis. We attempt to distinguish when possible between the requirements for the data storage system vs. the Web-based access to the data storage system. Overall, the purpose of this section is to provide the basis for evaluating a particular approach: that is, we assert that the degree of success by which a particular virtual organization data management architecture is "effective" is the degree to which that data management architecture satisfies these requirements. In particular, in Section 3, we present our approach and evaluate it based on these requirements in Section 2.

A. Holistic Requirements

In addition to being efficient and high-performance, a large-scale data management system for shared scientific datasets must exhibit the following properties:

Secure: While many scientific data sets are in freely-available and in the public domain (and hence do not require secure mechanisms and policies in place), most scientific data requires access control and accountability (e.g., to determine *post facto* who has accessed the data) for a variety of reasons. For example, even when the policy is such that anyone can access the data, often a person must first register and subsequently authenticate before access is granted (for example, so that the impact of the data might be attempted to be quantified via number of unique data accesses). The overall system must meet the collective security requirements (policy and mechanism) for the disparate collection of users.

Scalable: The system must be scalable along a number of dimensions: size of dataset managed, size of meta-data managed, and number of active participants (authors, curators, publishers, and consumers). Note: scalability does not necessarily require distributed management of distributed data. For example, the TerraServer [12] has 6TB in a SQL database, showing the effectiveness of single logically-centralized approach. In fact, for many large-scale science problems, economics argue that the data should be centralized and that computations should take place where the data is already resident [13].

Searchable: Consumers must be able to easily find the data they need, consumers need to be able to find the relevant metadata for their scientific explorations, authors must be able to easily find potential consumers of their data, etc. Collectively, users require to be able to search (and otherwise explore) both based on keywords and on application-specific properties of the data – e.g., "locate all scientific output (papers, derived datasets, etc.) directly or indirectly based on observations from Sensor21 in the range January 1 1985 through June 30 1988."

No special-purpose software: We strongly believe that the participants (authors, curators, publishers, and consumers) should not be required to learn new software packages in order to fully participate in the virtual organization.

Provenance: The data and metadata that is held by the data management system is connected via potentially complex set of relationships. For example, a potential consumer of a particular set of data might ask a question about the data in a particular blog, which might generate an answer that explicitly references another piece of data or metadata such as another blog entry. The data management system must be able to keep track of such histories and origins of data and metadata, and such provenance must be efficiently integrated into the rest of the data management system (such as search).

Notifications: The users of the system should not be expected to directly engage the data management system in order to determine what has changed since the last time they visited the system. That is, the users of the system should be able to register their interest in a variety of types of additions/modifications (e.g., data revisions/additions, metadata revisions/additions, new users of a particular class, etc.) and be able to receive these notifications via a variety of mechanisms (e.g., email, SMS, etc.) In essence, the system should selectively *push* information to the users of the system.

B. Requirements for Authors

The authors are the producers of the raw data. It is assumed that the authors have gathered/observed the raw data and performed some rudimentary processing on the data before they attempt to upload the data to the data management system. We make the assumption that there

are one or more curators for the raw data in question – the curator is not the same person as the author of the data, and the curator has the general responsibility of ensuring that data quality and correctness are achieved (in particular when compared across multiple raw data sources). In the case that there is not a specific curator, then the author is responsible for the curator's actions. The authors place the following requirements on the data management architecture:

Monitor data quality: Once the raw data is introduced to the system, the curator can modify/transform parts of the data to ensure its quality as compared to other raw data. The author requires an efficient mechanism by which to monitor their data as well as data products and summaries derived from their data and a means by which to record approval/disapproval of such modifications/additions. Note that the authors are expected to ensure the quality of the initial raw data, but this occurs prior to the raw data entering the data architecture system and as such places no specific requirements on the data architecture system.

Metadata: The authors must be able to easily view and provide metadata on the raw data, such as the attributes of the devices used to record such data, statements of confidence regarding the quality of the data, policies for using the raw data, properties of the measurement site, etc.

Request additional data: Often potential consumers of data will request additional data that is not present in the system currently. Ultimately, authors will need to be able to determine which requests fall under their purview.

C. Requirements for Curators

The curators ensure the quality of the author data largely through virtual organization-wide policies and/or via comparison with more than one data set. The raw data can contain gaps, errors, inconsistencies, the "wrong" units, etc., and it is the responsibility of the curator to properly address and/or "normalize" the data across multiple raw data sources. A virtual organization can have multiple curators who communicate and cooperate with each other to ensure the overall integrity and potential impact of the data. As such, the curators create a number of requirements for the data management system:

Existence of data/metadata: While consumers can make specific and directed requests for data to particular authors, it is also the role of the curator to interact with authors to obtain data and meta-data updates. The data management system must aid the curators in this respect by providing an easy means by which to determine which data is missing, who is responsible for providing it, when previous (unfilled) requests for the data have been made, etc. The data management system should facilitate such communication and tracking of requests.

Quality of data/metadata: The curator(s) develop standard quality-checking and processing algorithms and methodology that are domain-specific; the data management architecture must provide as much automation as possible by which to engage such functionality without manual (and often tedious) intervention by curators. In addition, users of the system can sporadically submit requested changes/clarifications to the raw data or the metadata. The data management system must provide a means by which the curator can easily review and approve/reject such requests.

Clarity of process: The data management system must provide mechanisms by which the curator can *explain* all actions – most notably, why/when a suggested modification to data/metadata was accepted/rejected. Note that not every instantiation of the data management system for a particular virtual organization is required to use such capabilities; rather, we believe that many virtual organizations will be more effective if such decisions are explained, and, as such, we believe that the data management system must be prepared to support such requirements.

D. Requirements for Publishers

In contrast to the authors, curators, and consumers – who clearly require domain knowledge to participate in the virtual organization – the publisher need not have an understanding of the raw data, the raw data format, the scientific hypotheses under consideration, etc. Rather, the publishers are similar to a traditional "system administrator" role, roughly responsible for keeping the system running. The publishers place the following requirements on the data management architecture:

Versioning and backup: While the data management system is continually evolving, the publishers require the ability to create major and minor versions of the entire system (data and metadata).

User creation, suspension, and termination: It must be easy to create new users, suspend accounts, and terminate users upon violation of virtual organization policy.

Availability of data/metadata: It is the publishers' responsibility to ensure that authorized users can access the data and meta-data for browsing, analysis, and download. If possible, the publisher should provide a means by which to aggregate data for efficiency, particularly given that access to information in the data management system will occur over the Internet.

Documentation: The publisher must ensure that virtual organization policies are readily discoverable. In addition, the publisher requires the ability to (along with the other roles) determine when such policies appear to be violated. The publisher also must be able to detail a clear path to becoming a data author or data consumer.

E. Requirements for Consumers

The consumers form scientific hypotheses and attempt to use the data and metadata as evidence for or against the particular hypothesis. The consumer can interact with all of the three other roles (author, curator, and publisher); many such requirements have been previously identified in this section. The addition requirements created by the consumer in the data management architecture are:

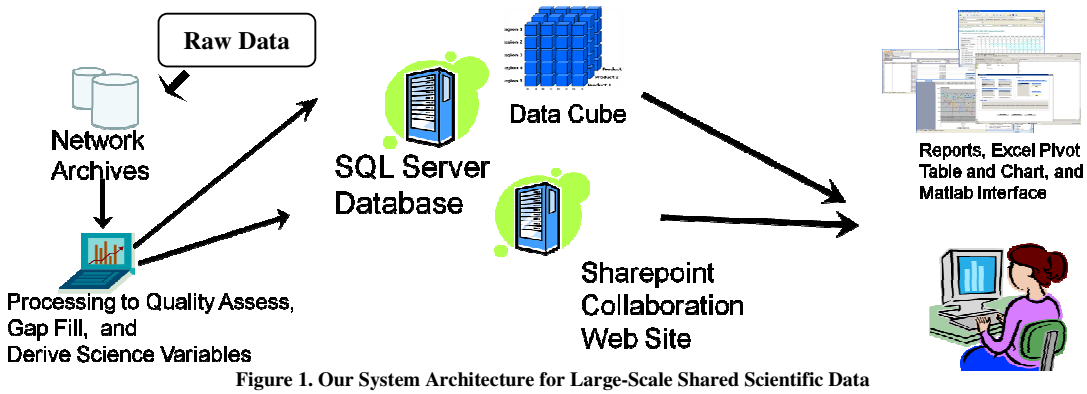

Figure 1. Our System Architecture for Large-Scale Shared Scientific Data

Request admission: Potential consumers wishing to use the data/metadata for analysis must be able, in effect, to request admission to the virtual organization. It must be possible to require additional specific information for the virtual organization such as their reason for requesting admission (what scientific hypothesis he/she will pursue), what data/metadata is intended to be used, whom the requester would collaborate with, etc.

Declaration of intent: Once a proposed consumer has been accepted into the virtual organization, the consumer must be able to record such interest and intent (as indicated in their application). Often, the data management system must supply support for making an explicit request to use a particular data (admission to the virtual organization can be decoupled from the actual use of data) and specify a means by which to receive notifications regard the update of relevant data/metadata.

Potentially multiple methods by which to access data/metadata: While there is a requirement that the data management system not impose special-purpose software tools in order to participate in the virtual organization, on the other hand, the virtual organization might have a set of mechanisms already existing by which the consumers access the data (e.g, FTP, HTTP, Matlab, etc.). The architecture should be flexible enough to support multiple mechanisms by which to access the data.

III. Our System Architecture

Having established requirements for each role as well as on the whole, we investigated existing options upon which to design and implement the data management architecture. We chose the combination of Microsoft SQL Server 2005 [14] as the back-end (centralized) data repository and Microsoft Office Sharepoint Server (MOSS) 2007 [15] as the underlying Web server platform. We chose SQL Server 2005 because of its long track record (although not for "scientific" data), scalability, 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 attempted to be used to meet the requirements of scientific collaborations.*

Our overall architecture and process is shown in Figure 1.Overall, the authors' raw data is first presented to the publisher often through a curator (as shown on the left of Figure 1). The publisher makes this data available back to the appropriate curator and author along with derived data products and data summaries (via "network archives") in order to enable additional quality control. This "second level curated" data is then given to the publisher to make available in three basic ways: direct access to the data stored in SQL Server, direct access to an "aggregate representation" (the "Data Cube" in Figure 1, described in Section 3.1), and access through the Sharepoint Collaboration Web Site (MOSS).

A. *Use of SQL Server and Associated Services*

In our approach, the raw data is stored in SQL Server, which has been shown to scale to extremely large datasets. The particular schemas are determined largely by the authors and curators. It is not strictly necessary for the consumers to know/understand the schemas if the particular consumer is only ever going to access the data through the Web interface or via the Cube interface (described below). Note that this does not imply that all data is assumed to be homogeneous – rather, there can be any number of schemas as defined by the virtual organization.

An important service provided in SQL Server 2005 is the *Analysis Services,* which provides support for Online Analytical Processing (OLAP). Over the past year, we've been experimenting using Analysis Services to build data cubes to support carbon-climate, hydrology, and other eco-scientists. Data cubes enable data mining and browsing. Simple aggregations (sum, min, or max) can be pre-computed. Additional calculations (e.g. median) can be computed dynamically or pre-computed. The data cube is constructed from a relational database using a specialized query language Multidimensional Expressions (MDX). The data typically can be organized along five major dimensions corresponding to: what (e.g. variables), when (e.g. time), where (e.g. geospatial location or site with elevation often included as well), which (e.g. versioning and other collection attributes), and how (e.g. gap-filling and other data quality assessments. Client tool integration is evolving, Excel PivotTables allow simple data viewing and we have enabled more

powerful analysis and plotting using Matlab and statistics software.

Computed members can be included in the data cube in addition to the usual count, sum, minimum and maximum. For example, hasDataRatio computes a fraction of data actually present across time and/or variables. Another example is DailyCalc and YearlyCalc which provide an average, sum or maximum depending on variable and includes unit conversion. YearlyCalc: similar to DailyCalc. RMS or sigma compute a standard deviation or variance for fast error or spread viewing. The computed members are typically d*riven by the nature of the analyses* – gaps, errors, conversions, and scientific variable derivations are facts of life for earth science data. Overall, we have found that data cubes provide a means of generating summary reports describing the data from a high level perspective which is essential to researchers looking for data to address a particular synthesis question.

B. Role-based Publication and Curation

We now describe the functionality of our Web-based server platform that leverages MOSS 2007. MOSS 2007 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 multiple different security models. MOSS 2007 adds search, basic collaboration, "business intelligence", and "enterprise content management". More specifically, MOSS 2007 adds RSS, blogs, wikis, etc. MOSS 2007 also has the ability to closely integrate with the Microsoft Office client suite (Excel, Word, etc.), although this is not specifically required or exploited in our architecture.

Many of the capabilities of MOSS 2007 readily map to the requirements that we established (and described in Section 2). We chose to utilize the MOSS 2007 ability to utilize Active Directory (AD) as an account manager and authentication source. Each user has a unique log-in and group membership(s) (e.g., author, curator, consumer, publisher). We leverage MOSS 2007's built-in ability to customize the server content based on ID and/or group to provide functionality based on role (e.g., only curators see "curator functionality"). Because each authenticated user automatically has a "Web space" in the MOSS 2007, we have, for example, the ability of each owner to place metadata regarding his/her data specifically on "their space", which is then searchable by the MOSS 2007 built-in search server.

In general, we have found that we need to create only a small number of specific functionalities (called "Web Parts" in the terminology of MOSS 2007) to address the requirements enumerated in Section 2. These Web parts, and the role(s) that see the capability, are:

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

IV. EVALUATION

We have performed a quantitative analysis of a number of routine operations for each of the four roles supported by our approach (author, curator, publisher, and consumer) for a representative test dataset and found that response time is sufficient, in particular as compared to latencies occurred in the wide-area (i.e., world-wide access to our dataset). As the dataset grows, some activities (particularly those involving AJAX) require real-time access to the back-end SQL database, which is only of minimal size and capacity in our test environment, and can incur a 1-2 second overhead. We believe that such latencies can be easily reduced by adding servers to the SQL server farm, but we have not directly done performed this test ourselves.

We believe that we meet the great majority of the "holistic" requirements in Section 2. In particular, security is ensured through a combination of browser-based username/passwords over SSL and Sharepoint's ability to restrict content/functionality based on Active Directory membership. All interactions with SQL Server are also logged. As mentioned above, although we do not have direct experience adding servers, we believe that there are sufficient capabilities in SQL Server and Sharepoint to readily increase capability to meet increased load requirements. Sharepoint features a built-in searching capability, to index based on keyword. We do not currently support non-keyword searching, which could be an important capability in the future. We rely on the browser as the sole required client-side software (Internet Explorer and Firefox are our focus). It is interesting to note that our main challenge with browsers has been debugging often idiosyncratic firewalls within enterprises that prevent direct access to our Sharepoint sites (we are slowly building a library of known firewall issues/configurations to better anticipate and debug future issues as more users register with our sites). Notifications are primarily provided by email and RSS. We only offer rudimentary support for provenance right now (e.g., blogs offer timestamps and hyperlink capabilities, and the SQL database has built-in support for versioning); we believe that better support for provenance is a critical capability that we must address in the near future.

Regarding the requirements of authors, we currently provide authors only limited support by which to monitor the quality of their data as a function of time. For example, an author has the ability at any time to download the current representation of their original data (perhaps modified via the curator's actions). Although the database contains a concise representation of the changes as a function of time, we have not had a need to expose this

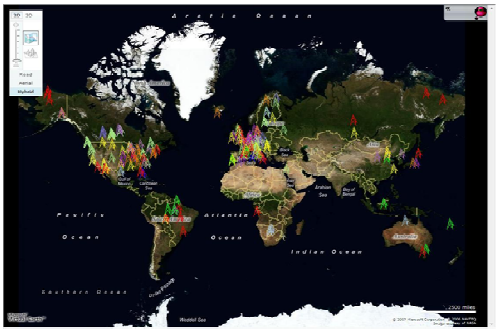
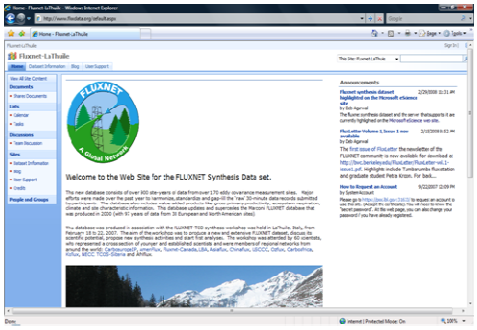

Figure 2. FLUXNET Synthesis Web Site (left: FluxData Towers; Right: Front page)

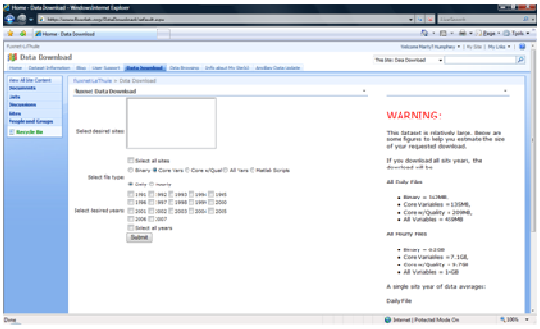
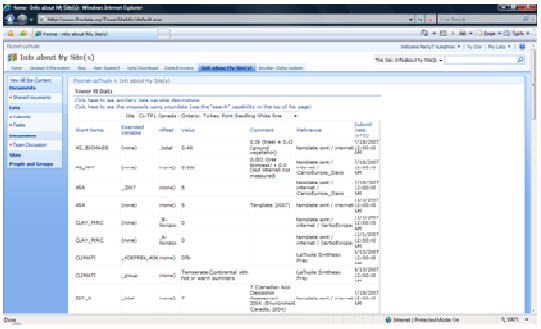

Figure 3. Subset of "Author" Support (left: "Download my data"; right: "Info about my site")

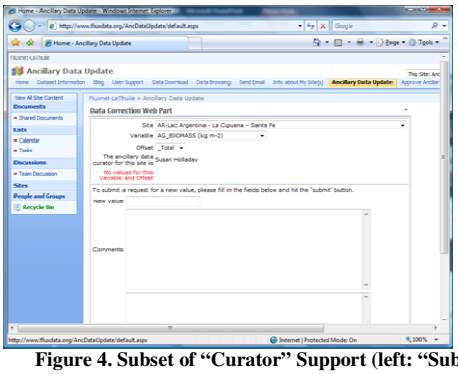
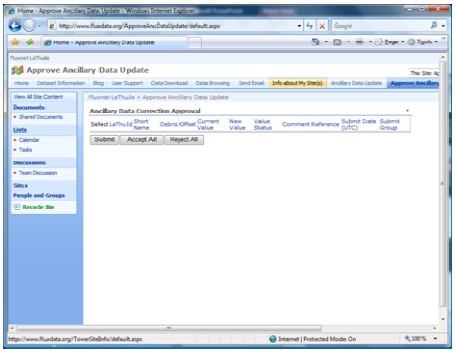

Figure 4. Subset of "Curator" Support (left: "Submit changes to ancillary data"; right: "Approve changes to ancillary data")

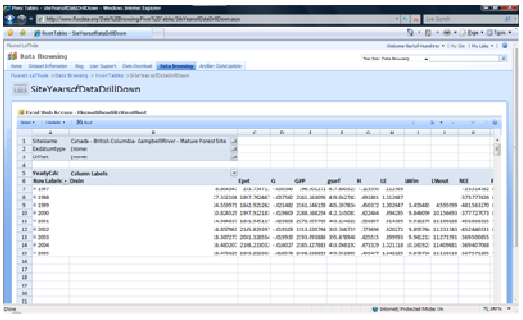
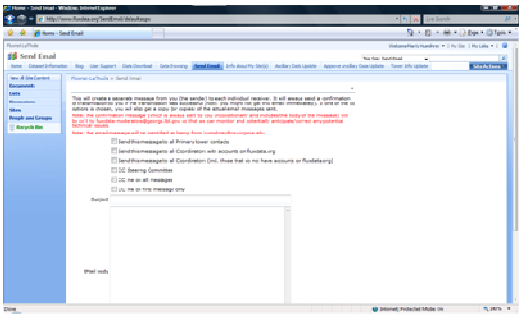

Figure 5. Subset of "Consumer" Support (left: "Visually inspect data cube via browser"; right: "Inform authors of data use and/or ask questions regarding the data")

representation directly to authors. An author has the ability to provide metadata, primary via their "My Space" within the Sharepoint server, which can then be found by others via a searching capability. A potential consumer can ask for "Additional Data" via a special email capability shown to them upon login, resulting in a directed email to the appropriate potential curator.

The curator requirement of receiving and monitoring requests for additional data is supported, although the subsequent "tracking" of requests is not directly supported in the system at this time. The requirement of the curator to assess the quality of the data is provided, albeit in a fairly primitive mechanism that can require more searching than can be desired. The main mechanism by which the curator can meet the requirement regarding "clarity of process" is through his/her blog or pages shared between curators.

Publishers have a requirement of "versioning and backup", which is provided intrinsically via SQL Server (because Sharepoint stores its data/web pages in SQL Server as well, the Sharepoint content is supported for versioning and backup). It is relatively easy to create new user accounts (in Active Directory). Regarding "availability of data/metadata", the security mechanisms of SQL Server and Sharepoint provide a robust access control framework, and the data cubes provide a means by which to aggregate the data for easier exploration. General support for Web pages and / or blogs provide the means by which the Publisher can clearly document activities and policies.

Consumers have the ability to "request admission" via an SSL-exposed page that asks the user to self-register (or submit an explicit request that must first be approved before admission is granted). The "declaration of intent" is not directly supported in our current system and is assumed to be outside of the system (this is part of the manual approval process to join the virtual organization). We currently support FTP and HTTP access to key datasets, as well as recent support for limited interaction via Matlab.

Overall, we believe that our combination of SQL Server, Sharepoint, and our 10 custom "Web Parts" meet the significant majority of requirements as established in Section 2, with the limitations described above. There are two significant areas that are not currently met, largely because of the complexity involved. First, we must ensure that the proper attribution occurs. For example, a long chain of information and/or events can be required before a scientific discovery takes place. Our architecture must ensure that this chain is easy to find, and is complete. Second, much of the processing in the system continues to rely too heavily on manual intervention. For example, while we have prototyped generic code by which to routinely build new data cubes, we have yet to deploy this on a routine basis. We plan to address both areas in the near future.

We have successfully applied our data management architecture as the basis for the global FLUXNET synthesis dataset collaboration. The FLUXNET synthesis dataset originally compiled for the La Thuile workshop in Feb 2007 contained approximately 600 site years. Over the year+ since the workshop, many additional site years have been added and the dataset now contains over 920 site years from over 240 sites. The ancillary data (metadata) describing the sites continues to evolve as well. There are on the order of 120 different data authors and 65 proposals submitted by teams of consumers to pursue synthesis activities have been approved to use the data. These proposals involve around 125 researchers. We have instantiated our architecture at http://www.fluxdata.org. The left side of Figure 2 shows a representation of the flux towers around the world that contribute data (in the role of "author"), and the right side shows the front page of the Web Portal. Figure 3 shows a subset of the support for the "Author" role for the FLUXNET project, Figure 4 shows a subset of the functionality for the "Curator" role, and Figure 5 shows a subset of the functionality for the "Consumer" role. Because of lack of space, only this subset is shown, and no functionality for the "Publisher" role is shown. Overall, through the feedback we have received from the FLUXNET participants, we believe that we are successfully meeting their requirements, and providing a high-quality and robust platform for dataset management.

V. CONCLUSION

Creating effective means by which scientists collaborate continues to be a significant challenge for today's Grids and eScience activities. By explicitly identifying and addressing the different requirements for each data role (author, publisher, curator, and consumer) in a large-scale virtual organization, we can create a data management architecture that enables the creation of datasets such as such synthesis datasets that continue to grow and evolve with new data, data annotations, participants, and use rules. We have evaluated our combined approach of SQL Server, Sharepoint, and our 10 custom web parts in light of these requirements and show how our data management approach is successfully being used for the FLUXNET synthesis dataset. In the coming months, we plan to migrate watershed data to the same infrastructure and add satellite and climate datasets to the global FLUXNET synthesis dataset. We plan to provide additional support for authors and consumers as identified earlier in this paper. In addition, we are currently in the process of making our software, including detailed documentation for its use, available for other projects.

REFERENCES

[1] B. Plale, D. Gannon, *et. al.* "CASA and LEAD: Adaptive Cyberinfrastructure for Real-Time Multiscale Weather Forecasting", Computer, Vol 39, issue 11, November 2006, pp. 56 – 64.

[2] US National Virtual Observatory (NVO). http://www.us-vo.org/

[3] Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydrologic Information System (HIS) http://www.cuahsi.org/his.html

[4] Biomedical Informatics Research Network (BIRN). http://www.nbirn.net/

[5] J. Gray, A. S. Szalay, A. Thakar, C. Stoughton, J. vandenBerg. Online Scientific Data Curation, Publication, and Archiving. MSR Tech Report MSR-TR-2002-74. July 2002.

[6] Y. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," in SIGMOD Record, vol. 34, 2005, pp. 31-36.

[7] A. Choudhary, M. Kandemir, J. No, G. Memik, X. Shen, W. Liao, H. Nagesh, S. More, V. Taylor, R. Thakur, and R. Stevens. Data management for large-scale scientific computations in high performance distributed systems. Cluster Computing. Volume 3, Number 1, July 2000, pp. 45-60.

[8] A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, and S. Tuecke. The Data Grid: Towards an Architecture for the Distributed Management and Analyses of Large Scientific Datasets.

[9] Nery dos Santos, M.; Cerqueira, R. GridFS: Targeting Data Sharing in Grid Environments, Sixth IEEE International Symposium on Cluster Computing and the Grid. Volume 2, Issue , 16-19 May 2006

[10] Storage Resource Broker (SRB): http://www.sdsc.edu/srb/index.php/Main_Page

[11] D. Baldocchi, Falge, E, Gu**, L.**, R. Olson, D. Hollinger, S. Running, P. Anthoni, Ch. Bernhofer, K. Davis, J. Fuentes, A. Goldstein, G. Katul, B. Law, X. Lee, Y. Malhi, T. Meyers, J.W. Munger, W. Oechel, K. Pilegaard, H.P. Schmid, R. Valentini, S. Verma, T. Vesala**,** K. Wilson and S. Wofsy. 2001. FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor and Energy Flux Densities. Bulletin of the American Meteorological Society 82: 2415-2435.

[12] TerraServer: http://www.terraserver.com/

[13] J. Gray, “Distributed Computing Economics”, Computer Systems Theory, Technology, and Applications, A Tribute to Roger Needham, A. Herbert and K. Sparck Jones eds., Springer, 2004, pp 93-101, also MSR-TR-2003-24, March 2003.

[14] Microsoft SQL Server 2005. http://www.microsoft.com/sql/default.mspx

[15] Microsoft Office Sharepoint Server (MOSS) 2007. http://www.microsoft.com/sharepoint/default.mspx