Publication and Curation of Large-Scale Shared Environmental 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
Microsoft Research Technical Report 93
July 2008

Abstract: *Many environmental scientists today need to assemble, use, share and save data from a diverse set of sources. These "synthesis" efforts are often interdisciplinary and blend data from ground-based sensors, satellites, field observations, and the literature. At even moderate scales of both data size and diversity, the cost and time required to find, gather, collate, normalize, and customize data in order to build a synthesis dataset can be daunting at best.. By explicitly identifying and addressing the different requirements for each data role (author, curator, data valet, publisher, and consumer), our data management architecture for large-scale shared environmental 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 Carbon-Climate Synthesis Dataset, one of the largest ongoing biogeophysical field experiments.*

1. 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. Recent progress in cyberinfrastructure has enabled easier, faster, and more secure access to these data sources and to the supercomputers needed to analyze the data. 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.

In the past, man environmental datasets were typically gathered, processed, and analyzed through a single heroic effort (often referred to as a *campaign).* The data is gathered, organized and processed to create science-ready *data products* through a single concentrated effort specifically for a set of known analyses by a set of known users. The data contributors know in advance exactly who will use their data, how their data will be processed, for what purpose their data will be used, and the agreed usage rules. This can work well when there are a relatively small number of data sources, there is little need for on-going data gathering or processing, and the data are generally similar in quality and form.

Data campaigns break down when a longer term *data occupation* is desired. Synthesis studies aimed at addressing larger or interdisciplinary questions are of increasing importance in environmental science. By definition, these studies bring together and utilize data from a diverse set of sources often at different time and length scales. The data may also span disciplines; for example, carbon-climate science brings together climatology, micro-meteorology, plant biology, and soil science. Synthesis can foster data reuse. Early results often raise new questions. The richness of synthesis data products can enable scientists from related disciplines. While wider use amortizes the initial

cost of producing the synthesis data, it can also lead to a need for ongoing data curation. Synthesis often requires a living dataset that can be cleaned, enhanced, and reused over time often in initially unanticipated ways.

In this paper we describe a novel architecture for the ongoing data publication, curation, and use of large-scale shared environmental data such as synthesis datasets. Our architecture is designed to enable environmental data to be collected, used, and maintained in a sustainable manner and should be 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.

Inherent in this architecture are a number of data roles. Each role has specific actions and responsibilities. The roles work together to facilitate the on-going and evolving data collection, processing, cleaning, and publication. In this work, we explicitly separate actions of each role in collecting, publishing, and using a dataset by a set of overlapping generic Web-based software capabilities for each role. Our data roles expand the author, curator, publisher, consumer taxonomy of [5]:

- An *author* produces the data and performs an initial quality assessment.
- A *curator* maintains the integrity of a collection of authors' data. The authors may be organized hierarchically (e.g., geographically), with a curator for each hierarchical group. Curators may also be responsible for a specific subset of the data across all authors. This is often the case when expert knowledge is needed for good judgment calls or when developing data processing algorithms or quality checks.
- A *data valet* deploys deploy curator-developed processing algorithms for deriving science variables from the authors' data. Working with the curators, the data valets clean and check the contributed data including regularizing formats and units.
- A *publisher* creates and operates the central starting point (e.g., Web portal) to search, curate, and obtain the data. The authors, curators, and data valets do not have the expertise, interest, and/or time necessary to provide the actual Internet-based access to the data. The publisher makes the data products produced by the data valets available to the consumers and enables the consumers to report issues with the data to the curators and data valets. An important contribution of the publisher is to understand and implement the data access policies that the authors/curators either collectively or individually impose.
- A *consumer* is the scientist pursuing an investigation that needs the data. The consumers are often organized into teams and work together on a topic.

The conventional approach to data management (e.g., in the Grid community) has been focused primarily on a computationally savvy scientist searching for relevant data from distributed sources [6][7]. The data are usually held in a file systems [8][9][10]. There has also been some important attention paid to meta-data creation and management (e.g., see [11] for a survey of data provenance systems). Unlike Grid community users, synthesis data authors and curators may or may not be computationally savvy. Moreover, the authors, curators, data valets, and consumers are not usually the same people. The author typically has expertise in the sensors and other devices for measuring environmental phenomena and in the particular domain science(s); the curator typically has expertise in the domain science(s). Both learn new software tools only when those new tools can simplify doing the science or enable new science. Today, the Grid data management tools rarely provide sufficient capabilities to easily inspect, correct, and otherwise update data while informing potential users of such data of such actions and why. As such, there is little incentive for non-computational scientists to learn the Grid tools.

We demonstrate the utility of our architecture in the context of the FLUXNET [12] Synthesis Dataset. FLUXNET is one of the largest ongoing biogeophysical experiments. The dataset contains data from 253 sites over ~960 total years of measurement. The data have been submitted by 140 authors and is being used by over 60 paper-writing teams of researchers worldwide. We developed and maintain the infrastructure and serve as the data valets and publishers.

The remainder of this paper is organized as follows. In Section 2, we establish the requirements, both on a per-role basis and for the system as a whole. Section 3 contains the details of our system architecture. In Section 4, we evaluate our approach and describe how we apply these principles to the FLUXNET Synthesis Dataset. Section 5 concludes.

2. Roles, Responsibilities and Actions

Surrounding an evolving shared dataset is a virtual organization. Within that virtual organization, there are a number of virtual relationships and on-going virtual conversations between individuals. Each individual may have one or more roles with actions and responsibilities. The virtual conversations between the roles are summarized in Figure 1 and will be discussed in the subsections to follow.

- The author can submit new data, download/inspect the data products derived from their own data and submit suggested updates or changes to their data or data products derived from their data.
- The curator can download/inspect submitted data, data products, or proposed data changes for which they are responsible. The curators also accept or reject all proposed changes.
- The data valets can create data releases, derived data products, data summaries, and accompanying documentation.
- The publishers can make new data products available, retire old data releases, change the user-interface aspects of access/control to

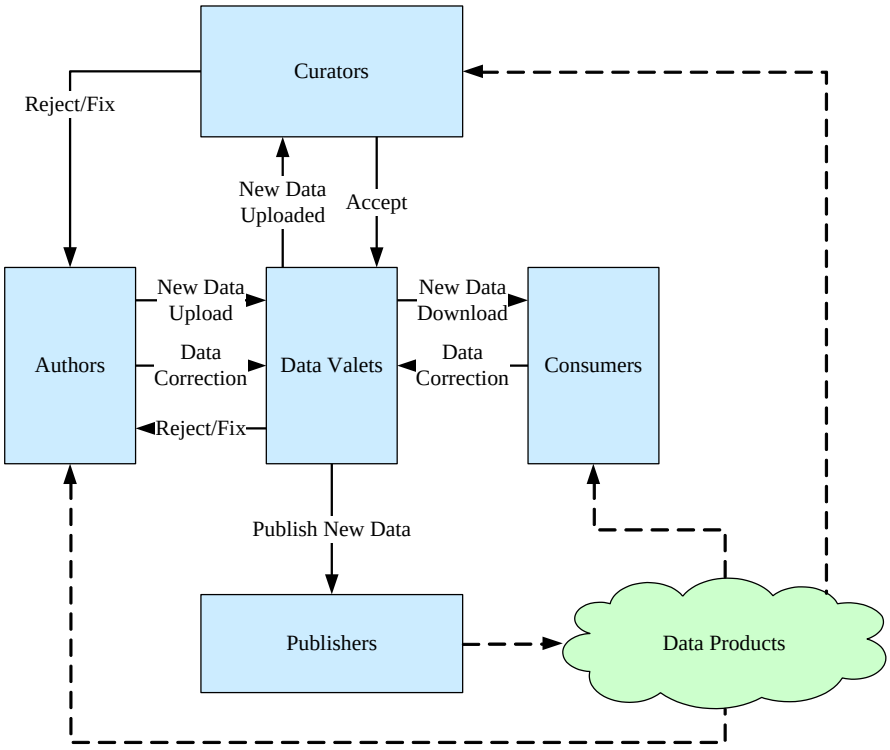

Figure 1. Data roles and interactions

datasets, and notify consumers of changes.

- The consumer can notify authors of data usage, communicate with the curators, data valets or authors to clarify data questions or suggest changes, and obtain author or other citation information for publication.

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. Some of these requirements overlap others, and some of these requirements conflict with each other. Note that 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, this section provides the basis for evaluating the approach we present in Section 3 or any other approach to virtual organization data management.

2.1 Authors

As the producers and contributors of the data, it is assumed that the authors have gathered/observed the data and may have performed processing on the data before they attempt to upload the data to the data management system. Typically, the author is interested in who is using his/her data and for what purposes. The author can provide insights on the origins of the data, data quality, and data applicability. Depending on the data sharing conventions, the author may also retain the ability to disallow the user of the data for a specific investigation. The author also wants proper attribution for the data; this may be crucial to ensure continuing funding for data collection. Therefore, authors place the following requirements on the data management architecture:

- ***Monitor data values and quality:*** Once the data is introduced to the system, the curators or data valets can modify/transform parts of the contributed data to ensure its quality as compared to other data. The data valets also produce data products using the contributed data. The author requires an efficient mechanism by which to discover changes to their data as well as access the data products and summaries derived from their data. The author also needs a means by which to record approval, disapproval, or questions on modifications and derivations**.**
- ***Contribute metadata and annotations:*** The authors must be able to easily view and provide metadata and annotations on the submitted data. This can include attributes of the instrumentation or methodology used to obtain the data, statements of confidence regarding the quality of the data, and suggestions for using the data.
- ***Contribute additional data or metadata:*** Often potential consumers 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.

2.2 Curators

A *curator* maintains the integrity of a collection of authors' data. 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. The authors may be organized hierarchically (e.g., geographically), with a curator for each hierarchical group. Curators may also be responsible for a specific subset of the data across all authors. Curators ensure the quality of the author data through data-set wide comparisons and analyses. Simple errors and inconsistencies are corrected by the data valets. The curator is responsible for corrections that need expert domain knowledge or mediation with the author. The curators create a number of requirements for the data management architecture:

- ***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, and when previous (unfilled) requests for the data have been made. The data management system should facilitate such communication and tracking of requests.

- ***Quality of data/metadata:*** The curator(s) define and may 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 or 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 curators can *explain* all actions. This is most important when a suggested modification to data/metadata was accepted/rejected. In our work with FLUXNET, we have observed that explaining such decisions reduces the number of repeated questions and concerns about the data. In other words, 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.

2.3 Data Valets

A *data valet* produces derived data products and, working with the curators, maintains the integrity of a collection of authors' data. Data valets form the first line of defense for the curator. The raw data can contain gaps, errors, inconsistencies, the "wrong" units, etc., and it is the responsibility of the data valet to properly address and/or "normalize" the data across multiple raw data sources prior to publishing submitted data to the curators. Data valets are usually organized by skill set and operate across all authors and curators. Unique requirements for the data management architecture by the data valets are:

- ***Algorithm development and deploymen:*** The data valets are responsible for the data processing algorithms and workflows necessary to produce data products. The algorithm may be specified by the curator, but ensuring that the deployment can be robust and "touch free" is the responsibility of the data valets. The data management system must include sufficient tools and import/export capabilities to support the data valets.
- ***Versioning and backup:*** While the data management system is generally viewed as continually evolving, the data valets require the ability to create major and minor versions of the entire system (data and metadata) when requested by the curators. A specific purpose of such versioning is to provide a dependable and well-documented view of the data for scientific analysis. Documentation of the contents of each version and the changes between versions must be readily available.

2.4 Publishers

The *publisher* creates and operates the central starting point (e.g., Web portal) to search, curate, and obtain the data. An important contribution of the publisher is to understand and implement the data access policies that the authors/curators either collectively or individually impose. The publishers develop and deploy the user interfaces used by all other roles and are responsible for keeping the system running. Usability is an important consideration for the 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 such knowledge. The publishers place the following requirements on the data management architecture:

- ***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 authors, curators, and consumers 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.

2.5 Consumers

Often organized into teams to work together on a topic, the *consumer* is the scientist who needs the data to pursue an investigation. The addition requirements created by the consumer in the data management architecture are:

- ***Request admission:*** Potential consumers wishing to use the data/metadata for analysis must be able 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 (e.g., what scientific hypothesis he/she will pursue). It may also be useful to be able to gather what data/metadata is intended to be used and with whom the requester would collaborate.
- ***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, as admission to the virtual organization can be decoupled from the actual use of data. The system should also 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 the data management system should not impose special-purpose software tools, the system should incorporate tools that are commonly used by the consumers. For example, if the consumers commonly use MatLab or NETCDF, native data access in those formats should be included. The architecture should be flexible enough to support multiple mechanisms by which to access the data.

2.6 Requirements for the Data Management System as a Whole

Like many data intensive systems, a large-scale data management system for shared environmental datasets must exhibit the following properties:

- ***Secure:*** While many scientific datasets are freely-available and in the public domain, most scientific data requires access control and accountability. The datasets may not require highly secure mechanisms and policies, but often must track or limit who has accessed the data. Registration and authentication may be used to quantify the number of unique data accesses or enable consumers to be notified of data changes. The overall system must meet the collective security requirements (policy and mechanism) for an often disparate collection of

authors, curators, data valets, publishers, and consumers.

- ***Scalable:*** The system must be scalable along a number of dimensions: number and sizes of datasets managed and number of active participants (authors, curators, data valets, publishers, and consumers). Note: scalability does not necessarily require distributed data. For example, the Sloan Digital Sky Survey [13] database now contains roughly 290 million objects and has 4 TB in a single SQL database, showing the effectiveness of single logically-centralized approach. 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 [12].
- ***Searchable:*** Consumers must be able to easily find the data they need with the relevant metadata, annotations, and commentary; authors must be able to easily find potential consumers of their data. Each role needs to be able to search (and otherwise explore) both based on keywords and on application-specific properties of the data. An ideal search might be "locate all scientific output (papers, derived datasets, etc.) directly or indirectly based on observations from the Sky Oaks site in the range January 1 1985 through June 30 1988."
- ***Ease of use for authors, curators, and consumers:*** We strongly believe that authors, curators and consumers should not be required to learn new software packages in order to fully participate in the virtual organization. The most attractive approach to this requirement is to ensure that these roles can fully participate using only their choice of Web browser.
- ***High-Performance:*** In addition to being stable and robust, the data management system must be efficient in all aspects of its behavior. This includes the ability to serve data to consumers, upload new data from authors, and show curators pending requested modifications.
- ***Provenance:*** The data and metadata that is held by the data management system are 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. This especially applies to the search capability.
- ***Notifications:*** The authors, curators, and consumers 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. Instead, these roles should be able to register interest in a variety of types of additions/modifications and be able to receive these notifications via a variety of mechanisms (e.g., email, SMS, etc.) Such registrations include not only data revisions and additions but also new consumers or new authors. In essence, the system should selectively *push* information to the users of the system.

3. System Architecture

To meet the requirements established in Section 2, we designed both a *data-centric* view of the architecture as well as a *collaboration-centric* view of the architecture. We begin this section with the data-centric view, as shown in Figure 2:

- The authors' data enter the data management system either directly or indirectly through a data archive. If indirectly, this archive ensures that the original data are preserved.
- The data enter the system in quarantine. The data may be published only to the authors, data valets and curators.

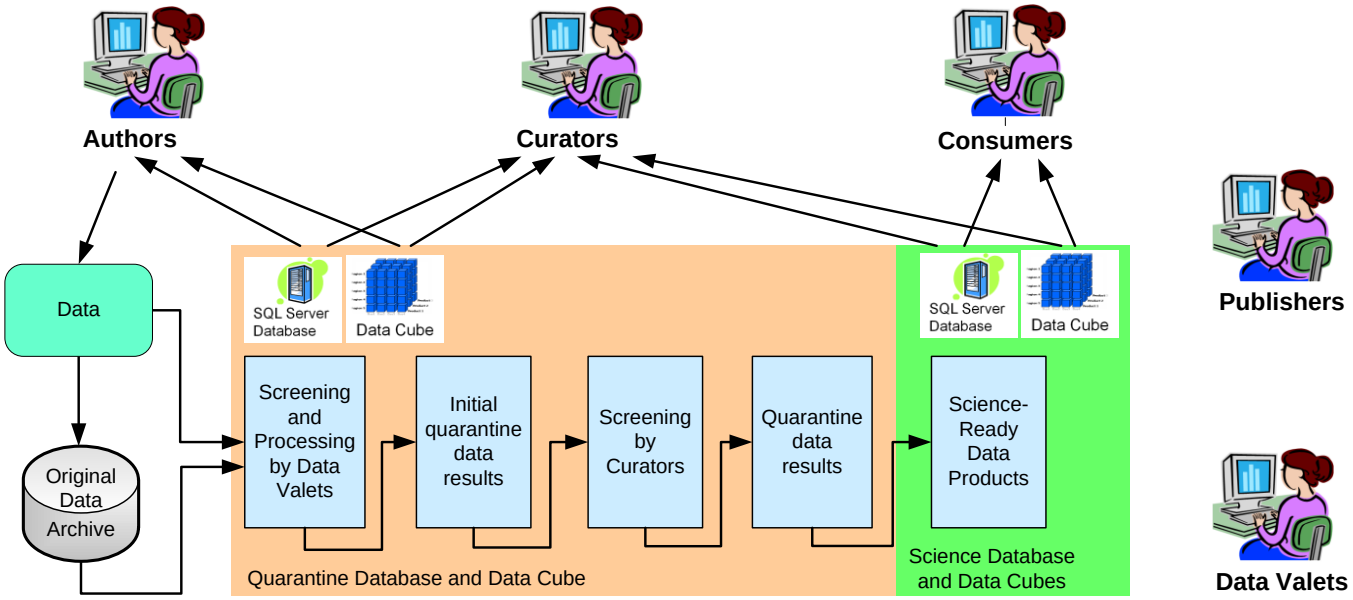

Figure 2. Data-Centric View of the Architecture for Large-Scale Shared Environmental Data

- The data valets perform the initial data cleaning and process the data for quality assessment.
- The data valets push the results of this processing to the publisher.
- The publisher makes the initial checking results available to the authors, data valets, and appropriate curator(s).
- The curators may perform additional data checks on the initially cleaned data. The curators push the results of these checks to the publisher.
- The publisher makes the additional curator assessments available to the authors, data valets, and appropriate curator(s).
- If there are any data quality or other issues with the data, the curators, data valets, and authors work together to correct the data. This may be done either by applying known and documented corrections to the initially submitted data or by the author resubmitting the corrected data.
- Once all data quality and other issues are addressed, the data is ready to leave quarantine. The data valets request that publisher retire access to the quarantine data as it is no longer necessary.
- The data valets perform additional processing to produce additional derived science-ready data products such as gap-filled files or data summaries.
- The data valets push the results of this processing to the publisher.
- The publisher makes these cleaned and curated data products visible to all parties in three ways: access to the data stored in one or more queryable data stores, access to summary data product files and data file download. While the data may be accessed in multiple ways, the queryable data store is the definitive store. All other data representations are derived from that store.

The collaboration-centric view of our architecture complements the data-centric view and is shown in Figure 3. Our collaboration Web portal contains:

- Shared content such as maps, blogs, data explanations, measurement site information, active papers and other information about the data and the data usage by consumers.
- Specific interfaces for each role enabling that role to have a virtual conversation with other roles.
- Protected and open access to summary reports which can also be downloaded.

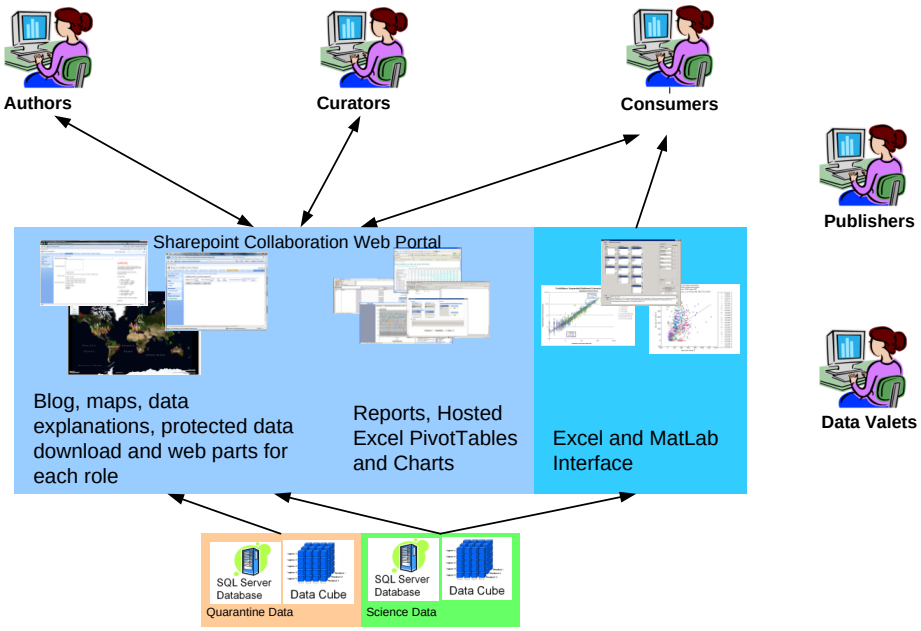

Figure 3. Collaboration-Centric View of the Architecture for Large-Scale Shared Environmental Data

- Protected data download to large data products such as gap-filled data files in MatLab or NETCDF format.
- Protected access to data browsing interfaces such as an Excel PivotTable. These interfaces directly connect to the queryable data stores.

In addition to the Web portal, consumers may access the database or data cube directly (subject to security controls).

3.1 Use of SQL Server and Associated Services

We chose the combination of Microsoft SQL Server 2005 [15] as the back-end (centralized) data repository and Microsoft Office Sharepoint Server (MOSS) 2007 [16] as the underlying Web server platform. We are targeting both Internet Explorer and Firefox to ensure wide applicability and interoperability for the client platforms. 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.

The particular database schemas are determined largely by the authors and curators. We have used a fully normalized schema; all data values occupy a unique data table row. This has the advantages of enabling us to add new and different variables to respond to changes in the dataset and to build data cubes directly from the database. The disadvantages are that the SQL query is less intuitive as the data are not organized in a simple spreadsheet like form with columns for each variable and the data table can have a large number of rows. Fortunately, it is not strictly necessary for the consumers to know/understand the database schemas. We export simple tabular views for some of the data and leverage data cubes for more general data browsing.

An important service provided in SQL Server 2005 is the *Analysis Services*, which provides support for Online Analytical Processing (OLAP). The data cube reduces the code the data valets would have to write to generate spatial or temporal

aggregations that feed into the summary reports and data products.

Over the past year, we've been experimenting using Analysis Services to build data cubes to support carbon-climate, hydrology, and other eco-scientists. While the science differs, the datasets have much in common. In particular, scientific data often naturally organizes along primary dimensions of what (variables), where (geospatial location or site), and when (time). Initially data cubes were developed for commercial needs like tracking sales of merchandise and financial data. Data cubes enable data mining and browsing. Simple aggregations (sum, min, or max) can be pre-computed. Additional calculations (e.g. variance or units conversions) can be computed dynamically. Hierarchies for simple filtering provide drilldown capability into each dimension.

A data cube is constructed from a relational database and can be queried using a specialized query language Multidimensional Expressions (MDX). We add which (versioning and other collection attributes), and how (gap-filling and other data quality assessments) to the what, where, when above. 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.

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. An example might be a researcher looking to understand the relationship between rainfall and carbon flux needs to be able to easily find sites with appropriate rainfall and vegetation characteristics.

3.2 Support for Role-based Publication and Curation

Having described how SQL Server and its Analysis Services create a scalable and efficient platform by which to store and access the raw data, 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.

MOSS 2007 is a sophisticated platform with a wide range of functionalities that readily mapped to the requirements that we established. 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, data valet, 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 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. Surface 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] , [Author], [Curator]

We have not yet had a need to create interfaces for the data valets. Through the combination of MOSS 2007 and custom Web parts, we are able to provide an interface that support publication of synthesis datasets and is easily maintainable.

4. Evaluation

We have informally performed a quantitative analysis of a number of routine operations for each of the five roles supported by our approach for a representative test dataset. We have found that response time is sufficient across most activities, in particular as compared to latencies occurred in the wide-area network and university networks used to 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 small 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 performed this test ourselves. In this section, we focus on the qualitative assessment of our architecture by determining the degree to which our

Figure 4. FluxData Synthesis Web site (left: FluxData Towers, right: Front page)

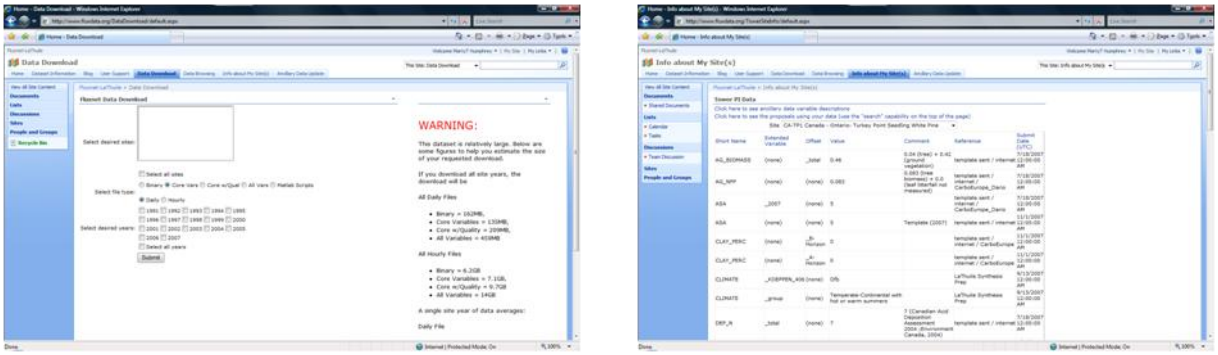

Figure 5. Subset of Author Support (left: “Download my data”; right: “Info about my site”)

approach is able to meet the requirements established in Section 2 for large-scale shared environmental data such as synthesis datasets.

We have successfully applied our data management architecture as the basis for the global FLUXNET synthesis dataset collaboration at http://www.fluxdata.org. The FLUXNET synthesis dataset originally compiled for the La Thuile workshop in Feb 2007 contained approximately 600 site years of time sensor data; each site year contains over 40 science variables. Over the year+ since the workshop, many additional site years have been added and the dataset now contains over 960 site years from over 250 sites. Another data refresh update is expected to increase those numbers in the next few months. The additional non-sensor field measurement data as well as metadata describing the site and sensor deployments 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.

The left side of Figure 4 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 5 shows a subset of the support for the author role for the FLUXNET project, Figure 6 shows a subset of the functionality for the curator role, and Figure 7 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 curator or publisher roles is shown. Overall, through the feedback we

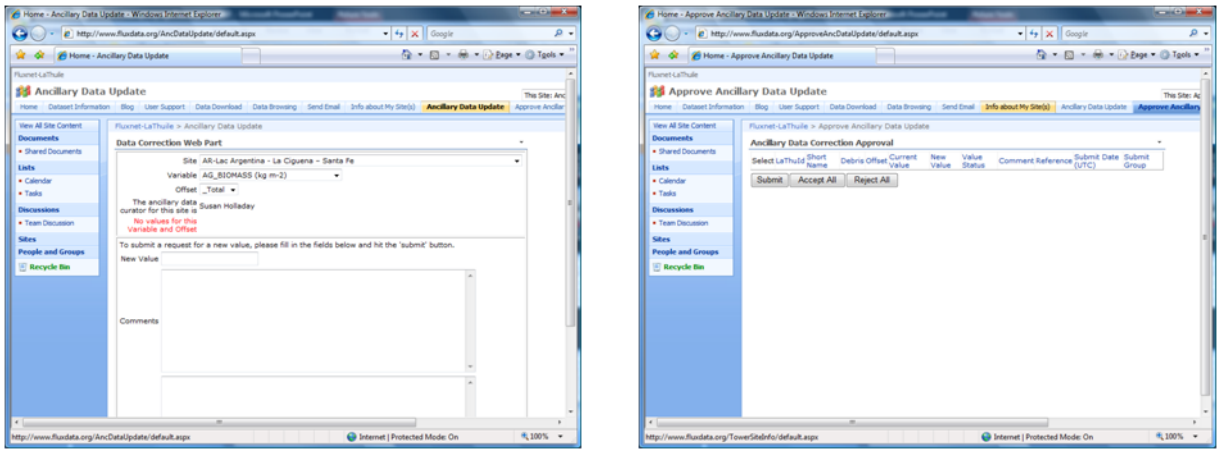

Figure 6. Subset of Curator Support (left: "Submit changes to ancillary data"; right: "Approve changes to ancillary data")

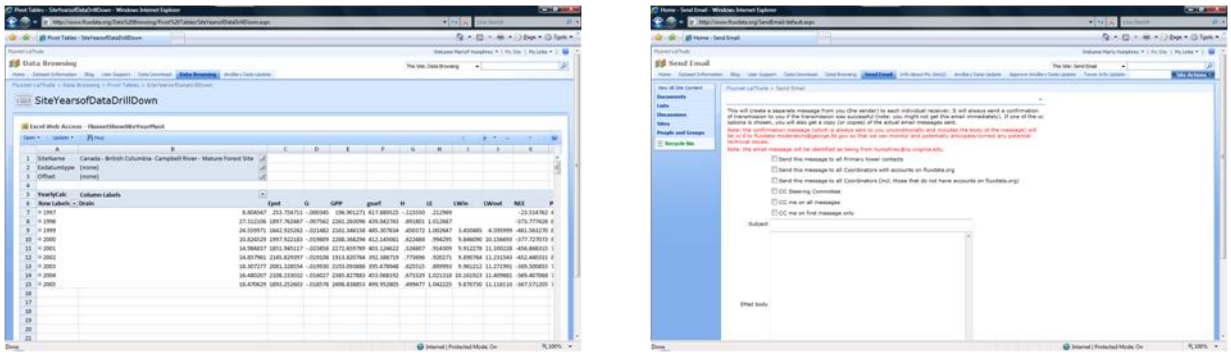

Figure 7. Subset of Consumer Support (left: "Visually inspect data cube via browser"; right: "Inform authors of data use and/or ask questions regarding the data")

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.

We believe that we meet the great majority of the "holistic" requirements in Section 2. In particular, the general requirements are addressed as follows:

- 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.
- Although we have not needed to add 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 for us to address in the future.
- We rely on the browser as the sole required client-side software (Internet Explorer and Firefox are our "official" supported platforms, although we routinely test via other browsers such as Opera). 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.

We currently provide authors only rudimentary 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 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 and/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), although this can be a somewhat tedious method if the number of groups the user belongs to is large. 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). This process works well in general, except in those situations where a firewall prevents the connection (often without explanation). 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 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 current 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. In general, much of the publishing can better benefit from automation. We plan to address both areas in the near future.

5. Conclusion

Creating effective means by which scientists collaborate continues to be a significant challenge for today's Grids and eScience activities. Arguably, in many situations, it is not sufficient to attempt to create a file system abstraction on distributed data and thereby believe that scientific discoveries will be significantly accelerated. In this paper, we have argued that by explicitly identifying and addressing the different requirements for each data role (author, publisher, data valet, 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 implemented and evaluated our combined approach of SQL Server, Sharepoint, and our 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 hydrological 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 to utilize.

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] 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.

[7] 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. Journal of Network and Computer Applications, 2001

[8] B. White, M. Walker, M. Humphrey, and A. Grimshaw. LegionFS: A Secure and Scalable File System Supporting Cross-Domain High-Performance Applications. In *Proceedings of Supercomputing 2001* , Denver, Colorado, November, 2001.

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

[12] 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.

[13] "The Sixth Data Release of the Sloan Digital Sky Survey", J. Adelman-McCarthy et al. 2007, The Astrophysical Journal Supplement Series, Volume 175, Issue 2, pp. 297-313.

[14] 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.

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

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