Scientific Data Management in a Web 2.0 Environment

By

Aaron Vimont

aaron.vimont@colorado.edu

A thesis submitted in partial satisfaction of the requirements for the degree
Bachelors of Science in Computer Science
Spring 2011

Advisor: Roger King, Professor, University of Colorado

Thesis Committee: Roger King, Professor, University of Colorado
Michael A. Eisenberg, Professor, University of Colorado
Qin Lv, Assistant Professor, University of Colorado

Table of Contents

Introduction

The modern world has become inundated with new technology. Computers are constantly getting faster, smaller and more capable than ever before. Information sharing has spread to a global community by way of the Internet. The Internet provides a way for anyone with a computer to share their ideas, thoughts and discoveries with millions of people across the globe. It transitioned from a tool where information can simply be viewed to a tool where information is meant to be shared. The common term for this transition is called the Web 2.0, where the user of a webpage has just as much influence as the creator. Web 2.0 absorbs nearly every aspect of the computerized world and drives the innovation and creation of thousands of new applications. Many uses of Web 2.0 applications fall in the realm of social enterprises. However, it also has begun to influence educational and scientific communities.

Cyberinfrastructure and e-Science arose as a way of merging scientific enterprises with Web 2.0 technologies and trends. Cyberinfrastructure and e-Science provide an entirely new set of tools for the scientific community, which should be adopted to improve the organization and distribution of scientific data. Many times, data collected and managed by members of the scientific community needs to be shared with dozens of other people, some of whom may be halfway around the world. This data helps shape the way the world is understood and perceived and can provide new methods of data collection and evaluation.

Web 2.0 technologies are often used to reach a broad audience. The webpages are designed so information can be shared with anyone in the world. This concept does not always directly apply to Cyberinfrastructure, which seeks to improve communications within a specific scientific community. I took the concept of Cyberinfrastructure and applied it to a lab at the University of Colorado. I developed a new system based on Web 2.0 technologies that uses a web interface and back-end database to store and manage data the lab collects. The data can then be shared with others from around the world. The old methods for storing and sharing data allowed for too high a margin of error. The project uses many aspects of Web 2.0 technologies, but to a limited degree. Only Invited customers and administrators can view the webpages developed for this project. The site is not available to anyone in the world like many Web 2.0 websites. Despite restricting user access, the Cyberinfrastructure helped ease the distribution of information and was not disruptive to the operation of the lab. In this paper, I explore the world of Cyberinfrastructure and e-Science by examining Web 2.0, applications of Web 2.0 technologies, the immense potential of Cyberinfrastructure, and a real world example of Cyberinfrastructure developed at the University of Colorado.

Web 2.0

The Internet moved away from static pages where the designer has complete control over the page. The user controls almost as much page content as the administrator. Dozens of websites now exist where content is entirely based on the user, such as Facebook. This world of user-based content has become known as Web 2.0. Interacting with information provided by others is now essential to almost all webpages. The Internet transitioned to a participatory, interactive place where information is created collaboratively [7]. The user comments, shares and posts this content, making it possible for breaking news to travel around the globe in a matter of minutes. Web 2.0 also makes it possible for a product to become an overnight success or instant failure.

Information now travels at split second speeds to almost anywhere in the world. Any successful web page, whether a social site or detailed representation of data, must be accessible to the user. Many times the success of these pages depends on user-to-user advertising [2]. A page with few visitors does not mean the page will be thrown out, but does often mean the implementation was a failure. Creating a web page for a known user base takes careful consideration of who will be using the site and what they will be contributing to the content.

Components of Web 2.0: Interface and Back-End

The most critical aspect of a Web application is the user interface. Before that can even be created, the designer must know whom the user will be. Without a known user in mind, the site will attempt to cater to everyone, but most likely fit the needs of no one. Like any product, time must be devoted to find the needs of the users. Time cannot be wasted with failed products and interfaces because companies use technology to become more efficient [7]. With a Web 2.0 application, a designer needs to understand some of the problems people are having with current technology, especially in an enterprise environment [13]. Recently, a survey showed that "many users felt current organizational communication/collaboration tools were not useful, with 15% of knowledge workers feeling they actually diminished productivity [13]." Designers need to focus on creating an application that can actually solve the issues in current technologies. Without a proper understanding of what the user needs, it will be very difficult for a designer to create a usable product.

A Web 2.0 application needs to have a powerful front-end interface. A powerful interface makes it possible for the user not only to contribute content, but it makes it easy for them to do so. "Web 2.0 applications use technologies which ease the creation and manipulation of content [2]." These web applications utilize tools that in many ways have become industry standards. Tools and languages such as API's, AJAX, Ruby on Rails, MySQL, JavaScript and others make it possible for the user to quickly add or update content [2]. The idea behind using these tools is to hide much of the functionality from the user. This makes the interface easier to interact with and add content to the site.

Back-End Database

With so much content being added to Web 2.0 applications, almost all of them require some sort of Database Management System (DMBS) to archive all the data. These systems have vastly improved over the past few decades. New database technologies can very efficiently process complex queries on enormous datasets [11]. Many of these technologies enable users to create, share, and analyze data in new ways [4]. Without these advances in database technology, many Web applications could not exist. Databases are complex, yet immensely powerful tools. The new advances make it possible to query the database in a more sophisticated way, which has now become an expectation of the user [11]. There is no longer a need for a command prompt. The user does not even have to understand SQL queries. Modern programming languages, like PHP, provide functionality specific to database interaction. This not only makes it easier for the user, but for the designer as well.

Modern DBMSs are still very difficult for most people to interact with and use. This is simply the inherent nature of databases, however this does not take away from the power these systems have. Due to this difficulty, a quality user interface makes it possible for everyone to utilize the functionality in the database [11]. The tools necessary for data management and integration in a Web 2.0 application must be usable by the general public [4]. A powerful and well-designed back-end database make it possible for Web 2.0 applications to function. The database is critical to success of a webpage. However, great engineering on the back-end can still be undone by a poor interface. A well-designed interface can often carry a product despite slow downs and imperfections on the back-end [5].

Many of the complexities of DBMSs lie within how queries are implemented and how the data is stored. The system itself needs to be easy to use. A simple DMBS makes it possible for a Web application to evolve and move as necessary. This includes making the database schema structured in an intelligent way. Also, many of the people involved with the creation of Web applications are not advanced programmers. It is not necessary for them to understand the inner complexities of the system [3]. The designers can then optimize the application by simply making small changes to DBMS. These systems have improved how we manage all of the data. A simple Web application with a back-end database does not need a technical support team. New tools allow designers to manage almost all of the system themselves.

Components of Web 2.0: Content

Perhaps the defining characteristic of Web 2.0 applications is the content that users can share [4]. The content varies immensely from site to site but user interaction with the content often remains the same. Having uniformity across most Web 2.0 applications makes it much easier to move from one application to another. The content of the application helps ensure that people will actually use the site. If they have a personal connection to content on the page, people will be inclined to continue using that service.

The content defines a website and is usually specific to certain users. In fact, one of the main reasons that someone initially visits a website is the content on that page [2]. When designing a Web 2.0 application, considering the types of content also becomes important. Many times a webpage contains much more than simple text. There are now ways to share audio, video, images, and other media. An application that facilitates this kind of modern content appeals to the user [12]. Content can also come from other applications on a personal computer, such as Excel. It may be necessary for an application to handle files and content that was not directly generated on that website. By including content useful to the user, Web 2.0 websites may bring in a wider range of clients than older static websites [12].

One advantage of Web 2.0 applications is the ability for some content to be updated in real time [8]. With many users examining and evaluating the content, mistakes can be found quickly and adjusted for. This is especially true in a professional environment. A Web 2.0 application can offer a centralized location for storing and updating important documents and other content [8]. Online applications can bring together people who would normally not have the opportunity to work with each other. These web applications help maximize the collective intelligence of the users through "dynamic information sharing and creation [2]."

Another benefit of storing and using content in a Web 2.0 environment comes from the adaptability of the application. Many web technologies make it possible for a designer to make small changes in the code that make larger alterations in the appearance and capabilities of the website. This makes it possible for the website to adjust to changing needs of the users and the organization responsible for application [13]. The content, driven by the users, defines what the webpage will be and often if it will be a success. Content is a critical component to every Web 2.0 application. It gives users a chance to interact with webpage and other people around the world.

Components of Web 2.0: User Profile

Users of a Web 2.0 application almost always have some form of user profile, which allows them access to the features of the site. Having a user profile has become almost synonymous with being online and using applications. Users of the page can then comment on, add and remove content. The profile helps give some sense of community to these online sites. Web 2.0 applications provide an innovative method of data management where more individuals contribute their own content [2]. The contributors do not need special knowledge of web programming or databases to make changes to the webpage. By using a much more user-friendly interface, web designers are likely to have more users on their site.

More and more activities are taking place on the Internet. The Internet is used for entertainment as well as business. New services are being put into an online environment [10]. Many of these online applications and new technologies have societal implications. They change the way people interact with each other and online services [10]. Web applications make it possible for people on opposite sides on the planet to easily interact. These online communities move online applications from isolation to interconnectedness for the programmers and the users [7].

One downside to online user profiles is the security risk. Peoples' profiles often contain lots of personal information. This can of course lead to security concerns, which are often not considered by the user [2]. Users are very comfortable with others viewing their profiles and often have very loose privacy settings. While loose privacy setting may be rather foolish, it does show the level of comfort people have in online communities. They feel they belong and can put personal information in their profile.

By having online communities driven by user profiles, it is possible to gain access to the collective intelligence of the user base. There are dozens of online forums where users can post questions and almost anyone in the world can answer. This is quick but can bring in concerns of authenticity [4]. Users should not rely on the accuracy of every forum. However, Web 2.0 applications make it possible to accumulate the "wisdom of the crowds [10]," and this can be applied to almost any field. While this is true, not every Web 2.0 environment is designed to be open to any user. Restricting the user base provides a means of quality control and allows more sensitive data to be place in these environments. Web 2.0 tools are used within scientific communities as a means of storing and sharing information. These tools are not used by the general public, but instead focus on a group or several groups of individuals.

Cyberinfrastructure and e-Science

Many Web 2.0 applications focus on entertainment and user opinions. Some of the largest and most popular websites in the world are completely based on the random thoughts and opinions of the users. However, Web 2.0 has moved into the world of scientific research and data collection known as Cyberinfrastructure and e-Science.

While this may not capture the interest of as wide of an audience, Cyberinfrastructure and e-Science can have a large impact on the modern scientific community. For all intents and purposes, Cyberinfrastructure and e-Science describe the same idea and therefore, Cyberinfrastructure will be used to describe both for the remainder of this paper.

Scientific collaboration itself is not simply a technological enterprise. It is primarily a societal and organizational issue [1]. Cyberinfrastructure has begun moving the scientific community away from desktop applications. Cyberinfrastructure describes the "infrastructure of distributed computer, information, and communication technologies supporting a transformation in sciences towards large scale, collaborative data driven enterprises [1]."

A Closer Look

Working in a scientific environment encompasses much more than simple data crunching. Many laboratories and other workshops must calibrate and build instruments specific for their research. Often the data collected by these instruments is meant to be shared with collaborators all over the world. Cyberinfrastructure provides a way for this collaboration to happen. As described by John Taylor, Director of General Research Councils UK Office of Science and Technology, "[Cyberinfrastructure] is about global collaboration in key areas of science [6]." It provides a system for virtual organization, which is supported by Web 2.0 tools. Cyberinfrastructure is designed to exploit Internet technology, such as programs or data sources built for distributed access [6]. The need to share information across the globe can easily utilize Web 2.0 technologies.

Web 2.0 can be immensely useful to Cyberinfrastructure. Web 2.0 tools can enhance scientific collaboration due to the support for virtual organizations [6]. Many current Web 2.0 applications already have a community of users. These users are familiar with the technology and because so many people are involved with these websites, there have been developments in high quality technologies and software.

Cummings et al describe what they call Web 2.0 proclivity, defined as someone's, "predisposition or propensity to use Web 2.0 tools [13]." Personal use of Web 2.0 tools will likely facilitate professional use (Figure 1) [13].

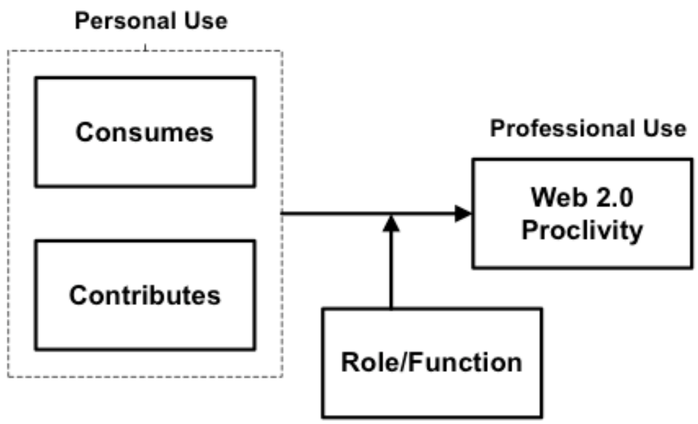

Figure 1: Model of Web 2.0 Proclivity

Web 2.0 technologies already focus on usability and user participation. Cyberinfrastructure can use this to bring science and its informatics to a broader audience [6].

In some ways it may be difficult to determine what Cyberinfrastructure should include because Web 2.0 technologies span dozens of audiences and applications. This will depend on the type of data that is to be shared but for a generalization, Cyberinfrastructure services fall into several categories. These include "models, applications, and simulations; data access, storage, federation, and discovery; filters for data mining and manipulation; and finally general capabilities such as collaboration, security etc. [6]." Fortunately, Web 2.0 technologies are already being employed to solve these very issues. They already include support for graphics, data storage, and data manipulation. Web 2.0 applications also make it possible to support a variety of desktop applications online.

Implementation

In the past, Cyberinfrastructure was implemented on a Grid system. Grids were, and still are, used to share information across the country and around the world.

However, Grids focused on secure, robust, and managed sharing of high value resources [6]. While this is an important part of data sharing, it does not invite interaction with the Grid. Grids require a more complex process to gain access and do not always provide the most user-friendly applications. Grids and Web 2.0 technologies actually address similar applications classes. Web 2.0 has focused more on user interaction, whereas Grids put emphasis on computing [6].

Web 2.0 technologies have been implemented on a Cloud system [9]. The Cloud is similar to the Grid, but Cloud architectures identify collections of services. Grid technologies identify individual services [6]. Cloud computing is more a set of pooled computing resources. In many ways Cloud computing fits the needs of Cyberinfrastructure better than Grids. Clouds are designed for collaboration. Information does not need to be stored on a personal computer because it can be accessed from anywhere with an Internet connection. The internal workings of the Cloud are often hidden from developers, who work with a simpler API [9]. Developing in a Cloud environment does not require a designer to know the complexities involved with the Cloud.

Using Cyberinfrastructure

Pace et al describe what they call an action-centered or practice-centered perspective of Cyberinfrastructure. The action-centered perspective is not just "information or data that is considered to be moved between people and devices. Instead, [it] suggests that artifacts... have deeper social and personal purposes in a shared collaboration space [1]." This is true for many aspects of Cyberinfrastructure and science in general. Scientific collaboration is not simply combining data and interpreting results. It often involves discussions about methods and reliability of data and how it will affect future research in that area.

Cyberinfrastructure allows for necessary interactions like these to occur more easily. Sharing data through Web 2.0 tools emphasizes solutions that encourage user control, creativity, and social action [1]. Shared data can then be interpreted and manipulated by those who have access to it. This may result in solutions to a problem or interpretations never before thought of by just one lab or group. Certainly not all data is designed for collaboration and sharing. Some data is sensitive or very valuable to those involved with the research. Cyberinfrastructure would not apply to such researchers. However, Cyberinfrastructure can still provide an easier means of collaborating with others in the field.

One of the biggest challenges facing Cyberinfrastructure is the actual adoption of its practices [1]. Many people in the scientific community have found a method that works for their data collection or sharing, and they are hesitant to give that up. Sometimes these methods are decades old, and the scientists and researchers have not even begun to look at new technologies. Because of this, Cyberinfrastructure needs to be eased into a current scientific environment. It would be immensely difficult to redesign the tried and tested methods of data collection already in place in a lab. Cyberinfrastructure can be used in conjunction with these older methods. It needs to emphasize a system that is in the "natural mode" with the data. Poorly designed systems are going to remain undesirable [1].

Cyberinfrastructure must focus on the ease of use with the interface, which is the same as a Web application. "Designers need to consider both interaction with the system and interaction between participants [1]." Improving the ease of use will make the system much more likely to be adopted. The scientists using these applications cannot simply be treated as outside users. They are participants with Cyberinfrastructure [9], which has the potential to redefine the relationship scientists have with their data. Providing more interaction with the data changes the process of their work and value of the results [1].

Cyberinfrastructure also requires the storage and sharing of large amounts of data. Therefore data management is a critical part of the success of Cyberinfrastructure. Sharing large datasets creates issues about data integrity. Additionally, "sharing data at larger institutional and organizational levels entails the broader social problems of determining and attributing intellectual property [1]." This problem is not easily solved. However, keeping, analyzing, and interpreting the data before it is shared on a global scale may help prevent these issues. Administering the accessibility and distribution of data falls upon the original owner. They cannot post data before they are ready to have it online. Accessing the data is a "critical point of entry into the scientific community [1]." Here again, Web 2.0 technologies can help. They are designed for easy sharing and communication of data. Often Web 2.0 data is simply in the form of a blog post or picture, but the concept remains the same. Developing Cyberinfrastructure in a way that supports outside access may help facilitate the use of these environments [1].

INSTAAR and The Stable Isotope Lab

Perhaps the best way to understand the role of Cyberinfrastructure is to look at an example of employing Cyberinfrastructure practices in a real scientific environment. I implemented a database at the Institute of Arctic and Alpine Research (INSTAAR) to improve their communication with customers and data management in general. INSTAAR, located in the East Campus Research Park at the University of Colorado, "develops scientific knowledge of physical and biogeochemical environmental processes at local, regional and global scales, and applies this knowledge to improve society's awareness and understanding of natural and anthropogenic environmental change [14]." INSTAAR focuses their research in high-latitude, alpine and other environments and has become a leader both nationally and internationally in these areas of study.

Included in INSTAAR's research are extensive field studies where researchers collect and analyze samples found in the harsh and variable alpine environments. The sample collecting falls into three main spheres of research: the Ecosystems Group, the Geophysics Group, and the Past Global Change Group [14]. The Past Global Change Group looks at recreating the dynamics of past climates and environments to help increase our understanding of earth and all of its components

[14]. One of the main labs within the Past Global Change Group is the Stable Isotope Lab (SIL).

The Stable Isotope Lab

The SIL uses stable isotopes to study and analyze the "biogeochemical processes that control environmental change on human timescales [15]." The lab uses state of the art technology for analyzing samples but does not take advantage of Cyberinfrastructure practices. The SIL works in collaboration with several research groups, including the Carbon Cycle Greenhouse Gases Group (CCGG). Working together, these groups collect and measure air samples gathered from around the world. One of the main samples sources in the SIL is the ice cores collected from sites in Greenland, Antarctica, high-altitude tropics (Ecuador, Peru, Tibet), and other regions [15]. The ice cores provide a rich archive of paleoclimate history. The ice cores are unique in that they have a combination of high resolution in their samples and provide a long time scale [15].

Through the use of isotope ratio mass spectrometers, the lab can examine a variety of samples including water (often from melted ice samples), air, carbonates, and organics. With these tools, the SIL gathers information on hydrogen, carbon, or oxygen isotopes found in the samples. The SIL is one of the few labs in the world with the capabilities and equipment to analyze these samples. Because of this, the SIL provides analysis for scientists and institutions around the globe [15].

Due to the amount of samples examined in the SIL, the lab keeps extensive records of the results. While the lab has state of the art equipment, it still did not have an efficient or effective method of storing and sharing all of the results it gathered from the samples. The raw data from the mass spectrometers was copied in an Excel spreadsheet, where macros and other functions were applied to the data. The finalized data was then copied into a larger spreadsheet with data from older samples. If the samples came from a scientist or researcher from another institution, the finalized data was copied into a customer log. Those results were then copied to another Excel spreadsheet and emailed to the customer. Additionally, there was no centralized location where the customer information, such as emails or addresses, was being stored. A lab employee relied on saved emails or stored email addresses to send important sample results back to a customer.

With state of the art instruments used to test samples from all around the world, it was a logical choice for the SIL to upgrade their means data management. The lab itself has almost no Cyberinfrastructure but absolutely has a need for it. By opening itself to Cyberinfrastructure, the SIL not only eases its own mean of accessing data, but also improves interaction with its customers. I took on the task of developing a DBMS for the SIL in order to help organize their results and customer information.

The Project

From the beginning, I wanted to design a system that was easy to implement but would still fit the needs of the Stable Isotope Lab. I originally sat down with the lab manager and several of the lab employees who had made some previous attempts at a lab database. From there, I established what would be the essentials of the database. The primary concern for the lab was moving the SIL customer log out of Excel and into a more stable environment like a database. The customer log contained the isotope results of samples sent to the lab from outside sources, as well as various bits and pieces of customer contact information.

Finding information on one customer proved to be difficult with the customer log being stored in this format. The entire spreadsheet, with several thousand rows, needed to be sorted to find any partially organized results for one customer. Sending the results to a customer involved copying rows from one Excel spreadsheet to another, and sending the new spreadsheet to the customer. Copying to another spreadsheet left significant room for error. The lab needed a more robust method of sharing the results with customers.

When the lab receives samples from customers, the samples often come packaged in small boxes. These boxes usually contain bottles or glass vials with the samples in them. Each bottle has a hand written label from the customer (Figure 2). The labels on the bottles correspond to sample names from the other lab. The SIL has a different numbering system for the same samples. Therefore, it fell on a lab employee to decipher the handwriting on a sample bottle, and match that sample name to a SIL number. If the original sample name was copied incorrectly into the customer log, the corresponding results could be easily mixed up.

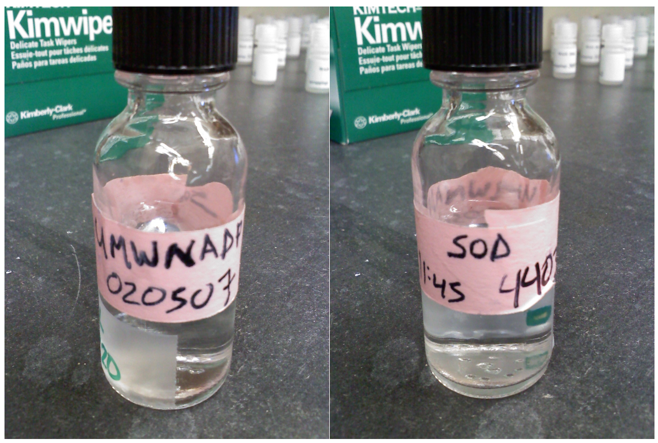

Figure 2: Sample bottle with complex label

There were cases when a customer could not match the sample names given back to them with sample names in their original list. The samples must then be re-run, wasting time and money.

Design of the New System

I took all of these factors into consideration when designing the new system for the SIL. The old methods of storing and sharing results were prone to errors that would be difficult to notice. I decided to implement a MySQL database with a web front-end interface. The interface would allow both customers and lab employees to modify the database easily and effectively without making accidental changes. This would also help the database be portable if the lab needed to use it locally or on the INSTAAR servers. I examined the current customer log and developed a schema that would breakdown the log into an effective database. For actual implementation, I researched several different methods. The first few methods I tried seemed to fit my solution; however, they proved too complicated or expensive to use.

Originally, I planned on building the database from the ground up. Using my own personal computer as a test, I downloaded files from the MySQL website to set up my computer as a server. Since I was using a Macintosh, the computer already came equipped with the Ruby on Rails. Ruby on Rails is a language designed for easy web development and interfacing with databases, making it a good candidate for the SIL project. I downloaded the NetBeans IDE to develop the Ruby code. Additionally, I downloaded the MySQL Workbench GUI toolkit, which is designed to ease the development of the database. The Workbench provides tools for designing the schema and queries without having to resort to the command line. Despite the supposed easy to install packages with MySQL, I found it immensely difficult to get the MySQL server running on my computer. When it finally was running, I could not find a way to connect to the server, even using the Workbench GUI. Therefore, the project had come to an unfortunate stand still. The installation and setup became too complicated, and using this approach would not allow for easy changes to the database. Because of these difficulties, I decided I needed a different approach.

To avoid complications of installing and managing a server specifically for the database, I researched other tools for database management. I found the program Filemaker Pro (FMP), which is ideal for small businesses and other similar projects. FMP takes almost all of the complications out of creating and manipulating the database. The interface is comparable to Excel. FMP also works well with Excel and similar dataset files. The SIL had a copy of FMP 10 installed on one of their Macintosh desktops. This gave me a chance to experiment with the program and see if it would fit the needs of the project. I found it easy to convert the customer log spreadsheet to the FMP database tables I had earlier designed. FMP improved interaction with the database itself, but it did not solve the problem of sharing customer data. The results would need to be copied into another document and emailed to the customer.

Using FMP Server Advanced, it is possible to build webpages that access the FMP server much in the same was a MySQL server. This would allow me to develop a small Web application where customers could see their results and upload sample information. However, FMP Server Advanced costs just under $3,000. While the SIL needed a database, this price was out of the question. FMP works well for small business applications, such as item inventory and budgets. For a project such as this, FMP could not provide the necessary tools.

Final Implementation

The solution I settled on used a combination of my previous two attempts. I liked the versatility of MySQL database, but installing proved too difficult. I also liked the features of FMP. It made it easy for me to change the database, but it did not provide the functionality necessary to put the database online. I decided to use MAMP Pro, a preconfigured server package, which helped avoid the complicated install and configuration of an individual MySQL server. MAMP stands for Macintosh, Apache, MySQL, and PHP. The downloadable package takes a few simple steps to install and provides a complete working server on a personal computer. Costing just $60, MAMP Pro is a much more affordable choice to FMP. MAMP Pro comes equipped with PHPMyAdmin, a web tool for easy modification of databases, tables, and data. PHPMyAdmin provides a GUI for managing the database (Figure 3). It allows the user to browse tables but makes it much more difficult to make accidental changes. MAMP Pro also utilizes PHP; so making webpages that interact with the database required mild PHP coding.

←T→			sampleID	silNum	externalID	analysis	sample	stdUsed	location
			1	11023	LFr-8m-12/28/05	O18	water	BW	Gandalf
			2	11024	LFr-9m-12/28/05	O18	water	BW	Gandalf
			3	11025	LFr-10m-12/28/05	O18	water	BW	Gandalf
			4	11026	LFr-11m-12/28/05	O18	water	BW	Gandalf
			5	11027	LFr-12m-12/28/05	O18	water	BW	Gandalf
			6	11028	LFr-13m-12/28/05	O18	water	BW	Gandalf
			7	11029	LFr-14m-12/28/05	O18	water	BW	Gandalf
			8	11030	LFr-15m-12/28/05	O18	water	BW	Gandalf

Figure 3: PHPMyAdmin interface for viewing tables

With MAMP Pro installed on a local machine in the SIL, I was able to move the customer log from the Excel spreadsheet into the MySQL database. PHPMyAdmin provides a GUI for uploading data from external files. I was able to take the Excel files and import them directly into the MySQL database with little trouble. From there I could locally develop Web applications for customers as well as lab employees before moving the database online. My original plan was to assign the local machine a static IP address, and customers could then access the webpages through that machine. One of the university server administrators prohibited this for security concerns. I was given permission to but the webpages up on the INSTAAR server, provided they had proper security scripts.

I used recommendations by Isaias et al while building my Web application. They detailed success factors for Web 2.0 applications, and I wanted my webpages to be practical for the users. To keep the database secure, I designed a page where customers would login before being taken to their homepage. I used PHP's md5() encryption function to encrypt the passwords of each user. The password retrieved from the login page is also encrypted and compared to the password in the database. Before any queries are run, the username and password variables are passed through PHP's addslahes() and mysql_real_escape_string() functions. These functions remove any characters from the variables that can be used for harmful SQL injections.

Once logged in, the user can view sample results, upload samples to be analyzed, change their customer information, or contact the SIL. The user can only view results and information associated with their username and ID. The homepage and all of the information on the page is specific to each customer. A list of sample results, organized by date, is dynamically generated upon navigating to the homepage (Figure 4).

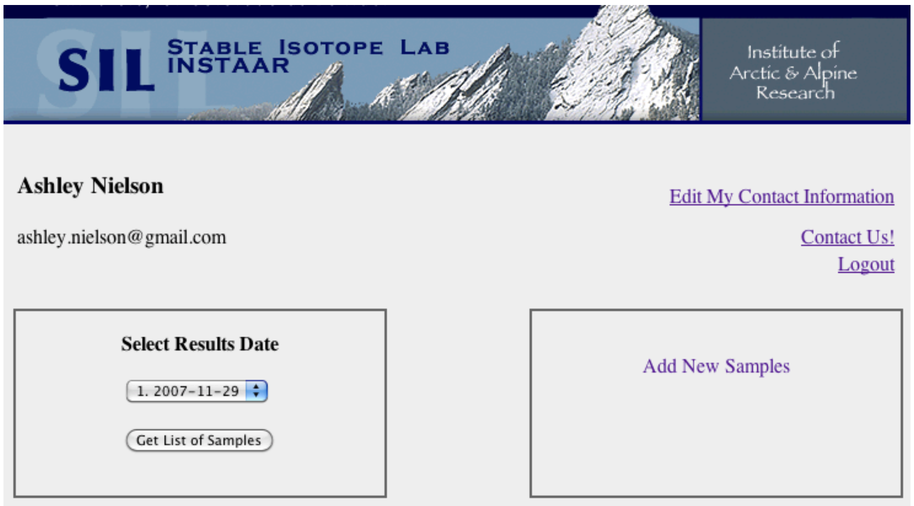

Figure 4: The site homepage designed for easy navigation

The user can choose which sample results to view based on a drop-down menu. After selecting which results to see, the user is taken to a page with a table of the results, generated through a small set of SQL queries (Figure 5). The table can be viewed in the webpage or downloaded as a CSV file to be used by other programs such as Excel. By putting the results online, it removes the hazard of copying and pasting Excel data. Not only can the customer see all of their results quickly and efficiently, but they could also download them and save them for future use.

```
SELECT SI.externalId, SI.analysis, ...
FROM sampleInfo SI WHERE SI.sampleID
      IN (SELECT SC.sampID FROM singleSampCustDate SC
      WHERE SC.rec = '$date' AND SC.custID
            IN (SELECT S.customerID FROM
            signIn S WHERE S.username = '$user'))
```

Figure 5: SQL query to build result table from database

The Web application also allows for the user to upload a new list of samples that will be run in the SIL. Many times, customers have dozens of samples they would like analyzed. The upload feature gives them the opportunity to take a list of samples they already have and upload it to the database. This prevents the lab from being responsible for copying in the names of the samples. By having customers upload information about future samples, the database is automatically updated for lab employees. They can see which samples will be coming in and once the samples have been processed, adding results becomes easy. It simply involves uploading a file with processed data, which is then added into the database (Figure 6).

```
while (($data = fgetcsv($fileDescript, 1000, ",")) != FALSE) {
     $sql = "INSERT INTO
     sampleInfo(silNum, externalID, analysis, sample)
     values('$max', '$data[0]', '$data[1]', '$data[2]')";
     mysqli_query($dbc, $import);
```

Figure 6: PHP code and SQL query to insert new customer samples into database

Contacting the customer after their samples have been analyzed is difficult when there is no central location where contact information is stored. Customers using the SIL Web application can follow a link to update not only their email, but also their institution, address and phone number. The application automatically fills in any information already present in the database, but gives the user the option to change it (Figure 7). This feature makes it much easier for the lab to get in touch with the customers to tell them their results are ready.

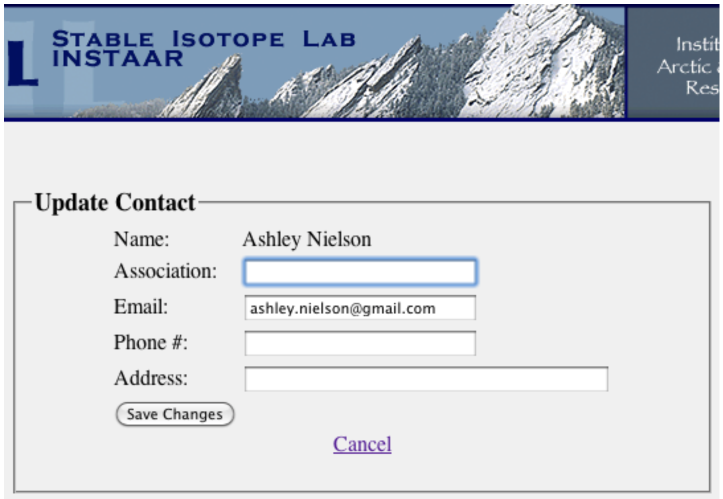

Figure 7: Editing customer contact information

It is no longer necessary to search through emails to find an appropriate return addresses. The homepage is simple enough where the user is not overwhelmed with options, yet it is personalized to them. They can see immediate changes they have made to their own contact information. Additionally, the "Contact Us" feature of the Web application makes it much easier for customers to get in touch with the lab should they have any concerns or comments.

I designed the SIL Web application to be simple to use while still providing the complex functionality necessary to update and change the database. The complexities are all on the back-end. The homepage minimizes the user options, but still allows them detailed interaction with their own page. The Web application moves the SIL into the world of Web 2.0 and Cyberinfrastructure. It did not require a major overhaul of their facilities or methods of data analysis. Using Cyberinfrastructure practices does not need to be an invasive procedure. Simply by taking the practices employed at the lab, I was able to create a more efficient and secure system for storing and sharing data.

While my system utilizes Web 2.0 tools, it does not encompass all aspects of Web 2.0. Many Web 2.0 applications are designed so any user can view the webpage and add content. For social sites, this may be appropriate; however, that was not appropriate for my site. The users need some measure of privacy for their information and results. Although the site does not support true Web 2.0 open information sharing, it still provides a structured and relatively open method of distributing scientific data and research. The design my website may in the future be used for implementing a system closer to true Web 2.0, such as encouraging input and comments from any member of the stable isotope community.

I created a system with a simple front-end web interface, making it easy for the user to modify and update the back-end database. I researched Web 2.0 and Cyberinfrastructure to build a better solution for the SIL. By combining Web 2.0 technologies with Cyberinfrastructure practices, my system avoids large disruptions to routines at the lab while improving the stability and accessibility of the data. Solutions such as the SIL Web application can be recreated in scientific communities around the world. Web applications provide simplicity, while expanding the audience it can reach.

Conclusion

The Internet drives the modern world. Web 2.0 applications make it possible to share information faster and in greater quantities than ever before. People now have a sense of online community where they can interact with others from across the globe. It is no longer necessary for the designer of the webpage to control all of the content. The user has the capability to add content and thoughts to a webpage. While many Web 2.0 applications are unique from one another, social functionality remains the driving force behind these webpages.

Cyberinfrastructure takes the functionality and popularity of Web 2.0 applications and applies it to scientific environments. It provides a new way for data and topics to be shared with scientists all around the world. Web 2.0 applications use powerful tools to store, share, and modify massive datasets. Using these tools for scientific research can greatly improve interactions with others in the same field.

The Stable Isotope Lab collaborates with dozens of other scientists and organizations. Their old methods for storing and sharing results provided too many opportunities for errors. By moving to a system that employs Cyberinfrastructure practices, they have not only made it easier to find important data but have also reduced errors. Cyberinfrastructure may not pertain to all scientific enterprises, but it also does not require a complete redesign of current methods. Cyberinfrastructure can easily be applied to current practices of a lab. The SIL took a simple problem of data management and was able to create a working system better for everyone involved. With the power of the Internet, creating Cyberinfrastructure shows immense promise for the future.

Works Cited

[1] Tyler Pace, Shaowen Bardzell, and Geoffrey Fox. 2010. Practice-centered e-science: a practice turn perspective on cyberinfrastructure design. In *Proceedings of the 16th ACM international conference on Supporting group work* (GROUP '10). ACM, New York, NY, USA, 293-302. DOI=10.1145/1880071.1880119 http://doi.acm.org/10.1145/1880071.1880119

[2] Pedro Isaias, Paula Miranda, and Sara Pifano. 2009. Critical Success Factors for Web 2.0 --- A Reference Framework. In *Proceedings of the 3d International Conference on Online Communities and Social Computing: Held as Part of HCI International 2009* (OCSC '09), A. Ant Ozok and Panayiotis Zaphiris (Eds.). Springer-Verlag, Berlin, Heidelberg, 354-363. DOI=10.1007/978-3-642-02774-1_39 http://dx.doi.org/10.1007/978-3-642-02774-1_39

[3] Julian Hyde. 2010. Data in flight. *Commun. ACM* 53, 1 (January 2010), 48-52. DOI=10.1145/1629175.1629195 http://doi.acm.org/10.1145/1629175.1629195

[4] Sihem Amer-Yahia, Volker Markl, Alon Halevy, AnHai Doan, Gustavo Alonso, Donald Kossmann, and Gerhard Weikum. 2008. Databases and Web 2.0 panel at VLDB 2007. *SIGMOD Rec.* 37, 1 (March 2008), 49-52. DOI=10.1145/1374780.1374794 http://doi.acm.org/10.1145/1374780.1374794

[5] John Canny. 2006. The Future of Human-Computer Interaction. *Queue* 4, 6 (July 2006), 24-32. DOI=10.1145/1147518.1147530 http://doi.acm.org/10.1145/1147518.1147530

[6] Geoffrey Fox and Marlon Pierce. 2009. Grids challenged by a Web 2.0 and multicore sandwich. *Concurr. Comput. : Pract. Exper.* 21, 3 (March 2009), 265-280. DOI=10.1002/cpe.v21:3 http://dx.doi.org/10.1002/cpe.v21:3

[7] Gwen Solomon and Lynne Schrum. 2007. *Web 2.0: New Tools, New Schools*. International Society for Technology in Education.

[8] Kera Z. Watkins. 2009. Peer evaluation as a needed web 2.0 activity in project management for teaching practical software engineering. In *Proceedings of the 10th ACM conference on SIG-information technology education* (SIGITE '09). ACM, New York, NY, USA, 173-177. DOI=10.1145/1631728.1631774 http://doi.acm.org/10.1145/1631728.1631774

[9] Marlon E. Pierce, Geoffrey C. Fox, Jong Y. Choi, Zhenhua Guo, Xiaoming Gao, and Yu Ma. 2009. Using Web 2.0 for scientific applications and scientific communities.

Concurr. Comput. : Pract. Exper. 21, 5 (April 2009), 583-603. DOI=10.1002/cpe.v21:5 http://dx.doi.org/10.1002/cpe.v21:5

[10] Johan van Wamelen and Dennis de Kool. 2008. Web 2.0: a basis for the second society?. In *Proceedings of the 2nd International Conference on Theory and Practice of Electronic Governance* (ICEGOV '08), Tomasz Janowski and Theresa A. Pardo (Eds.). ACM, New York, NY, USA, 349-354. DOI=10.1145/1509096.1509169 http://doi.acm.org/10.1145/1509096.1509169

[11] H. V. Jagadish, Adriane Chapman, Aaron Elkiss, Magesh Jayapandian, Yunyao Li, Arnab Nandi, and Cong Yu. 2007. Making database systems usable. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (SIGMOD '07). ACM, New York, NY, USA, 13-24. DOI=10.1145/1247480.1247483 http://doi.acm.org/10.1145/1247480.1247483

[12] Carsten Ullrich, Kerstin Borau, Heng Luo, Xiaohong Tan, Liping Shen, and Ruimin Shen. 2008. Why web 2.0 is good for learning and for research: principles and prototypes. In *Proceeding of the 17th international conference on World Wide Web* (WWW '08). ACM, New York, NY, USA, 705-714. DOI=10.1145/1367497.1367593 http://doi.acm.org/10.1145/1367497.1367593

[13] Jeff Cummings, Anne P. Massey, and V. Ramesh. 2009. Web 2.0 proclivity: understanding how personal use influences organizational adoption. In *Proceedings of the 27th ACM international conference on Design of communication* (SIGDOC '09). ACM, New York, NY, USA, 257-264. DOI=10.1145/1621995.1622047 http://doi.acm.org/10.1145/1621995.1622047

[14] "Introduction: Our Mission and Organization." INSTAAR, University of Colorado. 2003. Mar 8, 2011.

[15] "About SIL." INSTAAR, University of Colorado. 2005. Mar 8, 2011.