Connecting to the Data-Intensive Future of Scientific Research

Lafe G. Conner, Richard A. Gill, Rory O'Connor
Department of Biology
Brigham Young University
Provo, Utah 84602, USA

Abstract— **In recent years enormous amounts of digital data have become available to scientific researchers. This flood of data is transforming the way scientific research is conducted. Independent researchers are in serious need of tools that will help them managed and preserve the large volumes of data being created in their own labs. Data management will not only help researchers get or keep a handle on their data, it will also help them stay relevant and competitive in increasingly strict funding environments. This paper provides summaries of best practices and case studies of data management that relate to three common data management challenges – multitudinous sensor data, short-term data loss, and digital images. We use a combination of open system solutions such as HydroServer Lite, an open system database for time series data, and proprietary tools such as Adobe Photoshop Lightroom. Each lab may require its own unique suite of tools, but these are becoming numerous and readily available, making it easier to archive and share data with collaborators and to discover and integrate published data sets.**

Keywords—data management; sensor data, best practices, HydroServer Lite; data lifecycle

I. INTRODUCTION

The flood of digital data now available through networks of environmental sensors, satellite transmissions, and large data repositories is transforming ecological research. Data-intensive methods for discovering, integrating, and analyzing large stores of digital data have likewise transformed research in physics, astronomy, oceanography, hydrology, economics, earth and atmospheric science, and epidemiology [1]. New branches of scientific inquiry, such as genomics, would be impossible without the infrastructure and computational systems that support big data. This infrastructure is rapidly changing with the advent of cloud computing and distributed digital data bases. We are witnessing a revolution in the way data are used to make new discoveries that largely depends upon the ways that data are created, as well as the ways they are stored, retrieved, and shared [2].

The unique challenges that arise from conducting data-intensive science have given rise to fields that combine computer science and data management with traditional content disciplines. Examples of these new hybrid disciplines include bioinformatics, ecoinformatics, and hydroinformatics, which now have established journals, professional societies, and graduate and undergraduate majors. Data management is now among the most important aspects of scientific research and discovery, and we are finding that we cannot always rely on the data-management tools and techniques that served our predecessors [3].

Data management includes the planning, collection, assurance, description, preservation, discovery, integration, and analysis of data [4]. The goal of data management is to link data to the metadata that describe them and preserve both together over the long-term for the benefit of current and future research [5]. The fact that major funding agencies now require detailed plans that describe the types of data that will be collected, the standards used to assure and describe them, and their long-term preservation and sharing witness to the growing importance of good data management practices [6, 7]. Researchers increasingly are required as a condition of funding to publish their data along with the results and summaries of their research. This means that establishing and following the established data management standards for description and preservation may weigh just as much to the success of research endeavors as the actually collection and analysis of the data.

Very large experiments and observational networks (i.e. the Large Hadron Collider, the Sloan Digital Sky Survey, and Project Neptune) devote as much as one quarter to one half of their research budgets to software and personnel whose primary roles are to manage data [8]. For independent research labs, the data-intensive future could mean a widening gap between big science and the work conducted by an individual advising professor and her students. But, it does not have to.

Open-system solutions to data management, archiving, and publication may provide the means for independent researchers to remain relevant and competitive in the future of scientific research. Our own lab is undergoing a data management transformation that promises to make our data more useful to ourselves in the short-term and extend their life and usefulness in the long term. This transformation is possible because of software engineers who specialize in open-system solutions to the storage and management of time-series data.

Throughout the remainder of this paper we give a brief overview of the data lifecycle and the essentials of data management. Next we describe three common data-management challenges that we face in our lab and include the best practices that are recommended to address these challenges and the steps we are taking along those lines. We include descriptions of specific data management tools and practices as examples of solutions that we have found and are trying to implement. We recognize that there may be many other solutions and some may be better than the ones we are currently attempting. The data-management challenges we face in our lab – keeping up with the flood of sensor data; instilling standard practices for collection, version control, and description to prevent short-term data loss; and preserving digital images along with their metadata descriptions – are challenges that we expect to be common to many labs in many fields of science. Therefore, we hope these descriptions may be instructive and useful to a variety of researchers, even though the emphasis of our lab is ecosystem ecology.

II. Extending the Data Life Cycle

A. *The Traditional Data Life-Cycle*

In a traditional ecological research lab, the data life cycle begins when the researcher makes a preliminary observation or frames a specific question as a testable hypothesis. The researcher makes a plan to collect data that can help him answer the question or explain the causes that created the pattern he observed in nature. This plan leads the researcher to collect specific data, and involve students and collaborators in collecting other types of data as well. Each type of data is collected, assured for quality, and converted into the necessary standard units. The data are then integrated and analyzed and the results written, reviewed, and published (Fig. 1).

Afterwards, the students graduate and the researcher moves on to the next question. Over time, the data that they collected decrease in value, a phenomenon known as data entropy [5]. In two or three years, the lab computer is replaced and the data either remain on the old hard drive or are migrated to compact disks and placed in a drawer. While it is possible that a future researcher in the lab may come back to the data, it is more likely to be the end of the life cycle for those data.

We now recognize that this method of handling data is not sustainable, nor is it desirable. Data have the potential to be useful multiple times for answering a variety of different questions asked by researchers now and in the future. And perhaps we should think of ourselves as data stewards, rather than data owners, especially if our science is funded through federal programs or by public institutions.

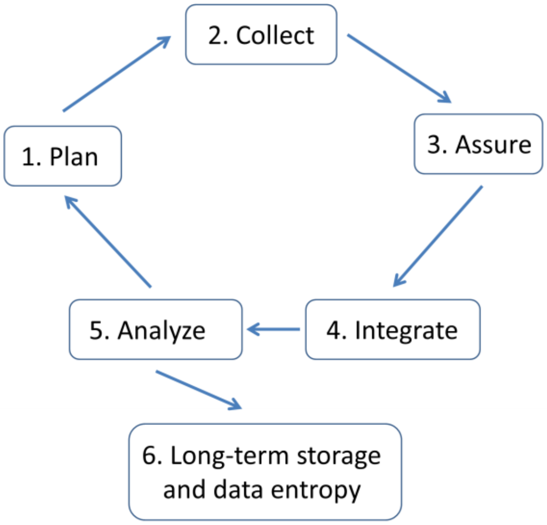


Figure 1. The traditional data lifecycle. Modified after Figure 1 in [4].

B. *Extending the data-lifecycle*

The intent of data-management planning is to consider the whole life of the data and to extend the traditional data lifecycle by encouraging researchers to describe, preserve, and share their data, so that the data may be discovered and integrated into new analyses (Fig. 2). A data management plan helps researchers consider and provide for the extended life of their data before they collect a single measurement. By knowing which types of data they will collect and where the data will be deposited for long-term storage, researchers can make use of established data management and data description standards. Data standards prescribe such things as the units and file types that data may be stored in and the metadata and specific vocabularies that can be used to describe them. There are even open source software programs that can be used to create metadata that meet specific standards [5].

Creating a data management plan has also become easier because of the Data Management Plan Tool (DMPTool) provided by the University of California Curation Center of the California Digital Library[1]. The DMPTool breaks down the writing of data management plans into sections by requirements specific to institutions and funding agencies. It guides researchers through writing a data management plan by asking specific questions about the data they plan to collect, the data standards they plan to employ, and the resources they will use to archive and share their data.

[1] https://dmp.cdlib.org/

Research described in this document is supported by fellowships from the Rocky Mountain NASA Space Grant Consortium Graduate Fellowship 2012-2103, the BYU Graduate Studies Fellowship 2011-2012, and the 2010 G.A. Harris Graduate Research Instrument Fellowship from Decagon Devices, Inc.

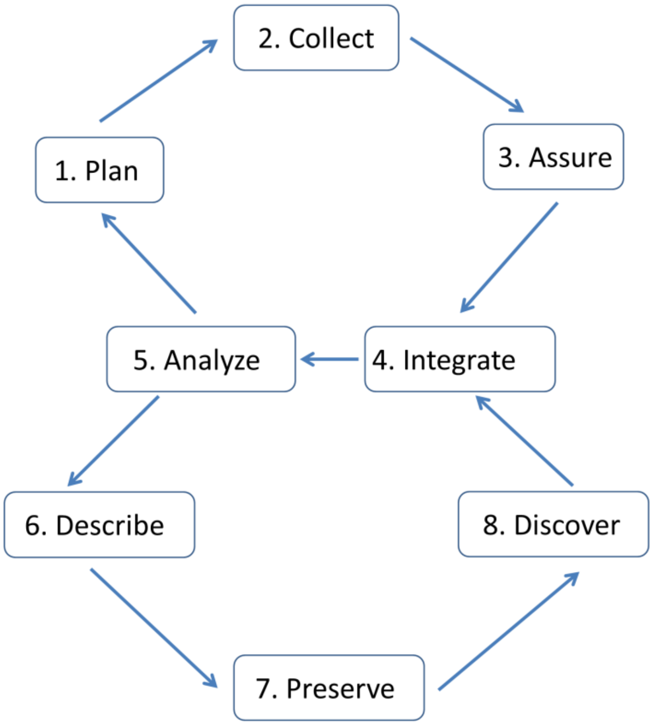


Figure 2. The extended data life cycle. Modified after Figure 1 in [4].

Automating data management processes helps to prevent errors from being introduced into the data and provides documentation of data management. Automated data management tools can assist in planning and documenting the use of data. Some examples include the automated workflow system Kepler,[2] which is similar to the model builder tool in ESRI's ArcGIS software (Fig. 3) [9,10].

Workflow tools automate and document data management tasks by identifying each step that data pass through as they move from raw formats toward the finished analysis. Workflow tools are linked directly to computing programs, so they actually perform the steps indicated in the workflow. Data management may also be documented by using automated scripts in programs such as Python and R [11,12]. Scripts and workflows can also be shared with collaborators, reviewers, and future data users, thus helping them to identify possible sources of errors and quickly modify specific steps in the process [5].

C. Opportunities for data preservation and discovery

There are a growing number of data repositories available to researchers. Repositories range from institutional archives managed by the library or special collections of a particular university to multinational collaborations such as the Global Biodiversity Information Facility (GBIF)[3] and the Knowledge Network for Biocomplexity (KNB)[4]. These repositories make it easier to share and to discover relevant data to a variety of research questions and, as data become more widely available, we may see a decline in the impact of studies that do not incorporate outside data that were discovered and integrated into the final analyses.

III. Case Study: Data Management Transforms the Gill Ecosystem and Global Change Ecology Lab

The Gill Ecosystem and Global Change Ecology Lab is an independent research lab at Brigham Young University. We study ecological responses to climate change. In recent years our research has focused on plant and soil responses to changes in water availability associated with climate warming and altered precipitation patterns. Some of the major data management challenges we face include (1) an overwhelming amount of sensor data from multiple datalogger platforms and experiments, (2) short-term data loss associated with frequent turnover in lab personnel, (3) data confusion due to multiple workbook and database versions, (4) from a lack of metadata sufficient to understand existing data, and (5) the organization, storage, and sharing of non numerical data, particularly digital images. For each of these challenges we describe the suggestions offered by sources like the Data Observation Network for Earth (DataONE), and tools that we are using and learning to use to achieve a higher level of data management.

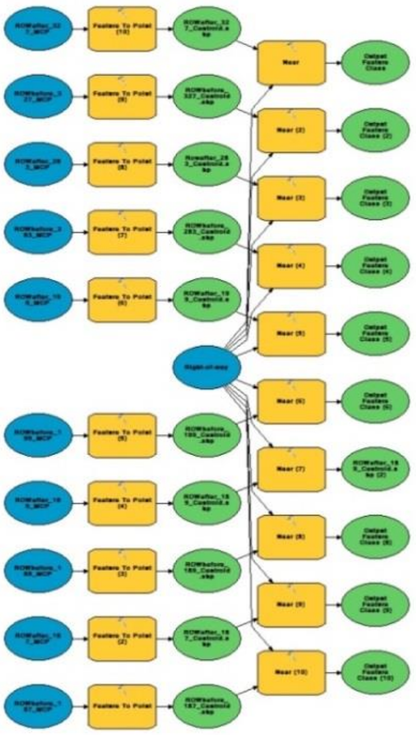


Figure 3. Sample work flow designed using model builder in ArcMap version 10. The blue ovals represent data input files, the yellow rectangles are operations performed on the data, and the green ovals are output files.

A. Challenge 1: Data Deluge in a Sensors Everywhere World

In the Gill Lab we are continuously collecting data from environmental sensors at sites across the state of Utah. These

[2] https://kepler-project.org/
[3] http://www.gbif.org
[4] https://knb.ecoinformatics.org/index.jsp

data are part of multiple experiments that run simultaneously and involve graduate and undergraduate researchers as well as collaboration with other faculty and their respective labs. For example, in a single experiment that has been running for three years, we are collecting data from sensors at thirty-two different sites. The sensors at each site record soil temperature, volumetric water content, and electrical conductivity averaged over six-hour intervals. In those three years we have collected more than 100 thousand data values that are stored in nearly 100 different files. We are capturing similar amounts of data from three other experiments and each year we add more sensors to new sites and new types of sensors to the old sites we already have. In this data environment, finding the right tools and having a working data-management system easily means the difference between having a successful research program and becoming obsolete [13].

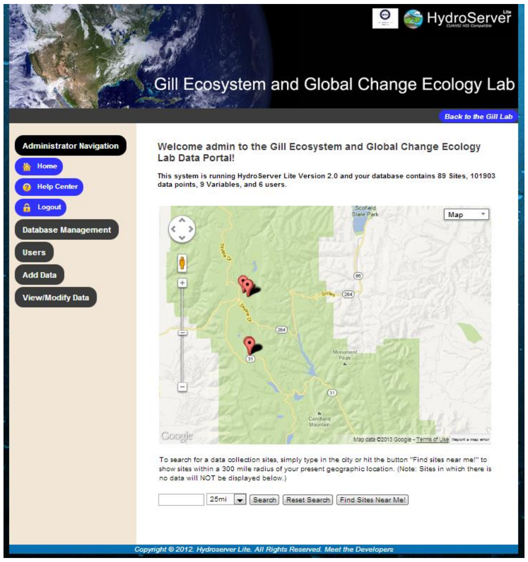
When data are retrieved from an environmental sensor they typically arrive as in a flat file (a text file or spreadsheet). The goal is to transfer these data from the multiple, disconnected files to a single relational database that will store the data values in a format that supports queries and retrieval. An example of the type of analysis we want to perform with these sensor data is to call up and compare volumetric water content in soils at different sites and different depths. However, when taken directly from the sensors the data values we want to compare are in separate spreadsheets between which there is no communication. The solution to this data management challenge is to store the data values in a relational database.

B. Best practices for managing sensor data

Ideally, systems for managing sensor data are automated to reduce human introduced errors and make the process quick, repeatable, and transparent [5,13]. Best practices require an establish system for conducting basic quality control and quality assurance as the data are taken from the multiple flat files and entered into the relational database, but that the data values going into the database remain as raw as possible [14]. If there are gaps to fill or corrections or manipulations to be made on the data before they are integrated and analyzed, then it is best to use a scripted language and save the script along with the metadata that describe the file. When filling in a gap by estimating a data value in a time-series of sensor data, it is best to identify which data are actual measurements and which are estimated [15]. After data have been entered into the relational database, they should be graphed to make sure that the data values match the metadata that describe them [16].

C. Hydroserver Lite as an Open-System Relational Database

To manage the flood of sensor data in our lab we needed a relational databases system, but it was beyond the training of any of our lab personnel to build one that would meet our lab needs and also comply with the standards that allow for data sharing within the discipline. Fortunately, we learned of the HydroServer Lite (HSL) database system. HSL is open-system and web-based, which means that an HSL database can be hosted on any webhosting service and accessed from any computer [17]. It also means the system is free and realtively easy to install.



Welcome admin to the Gill Ecosystem and Global Change Ecology Lab Data Portal!

This system is running HydroServer Lite Version 2.0 and your database contains 89 Sites, 101903 data points, 9 Variables, and 6 users.

To search for a data collection sites, simply type in the city or hit the button "Find sites near me!" to show sites within a 300 mile radius of your present geographic location. (Note: Sites in which there is no data will NOT be displayed below.)

Copyright © 2012. Hydroserver Lite. All Rights Reserved. Meet the Developers

Figure 4. Screen shot of the Gill Lab HydroServer Lite web-based relational database for storing time series data from environmental sensors.

With the help of Dan Ames we established an HSL relational database for the Gill lab in a single afternoon and began entering metadata and uploading data values within days.[5] HSL has a user interface that allows the administrator to create other users and define the permissions of those users. HSL users enter metadata through a similar user interface that includes controlled vocabulary for describing the data that align with the standards of the Consortium of Universities for the Advancement of Hydrologic Sciences, Inc. (CUAHSI). The relational database conforms to the observations database model (ODM) that is standard among hydrologists and that allows the data stored in an HSL database to be shared directly with other HydroServer and HSL databases. HSL is easily customizable and includes an interactive map feature that enables geographic searches and data selection (Fig. 4).

HSL contains features that allow for visualization of data as they are uploaded, which is an important feature for quality assurance (Fig.5). The web-user interface allows for making limited changes to data in the database. The database is built using MySQL, and direct connections to the MySQL database allow for more extensive querying and editing. For example, if data are associated with the wrong metadata a direct connection to the database can be used to erase the data values so they can be re-entered and associated with the correct metadata.

Database querying and data retrieval are accomplished using the open-system tool HydroDesktop, which includes applications for visualizing and editing the data and importing and analyzing the data through the HydroR connection to R [18]. HydroDesktop can access the HSL database through a direct connection but can also access the database through the CUAHSI Hydrologic Information System central registry.[6] This means that we can direct access to our data but so can our collaborators and anyone else who might find them useful.

[5] http://gilllabbyu.net63.net/client/

[6] http://his.cuahsi.org/

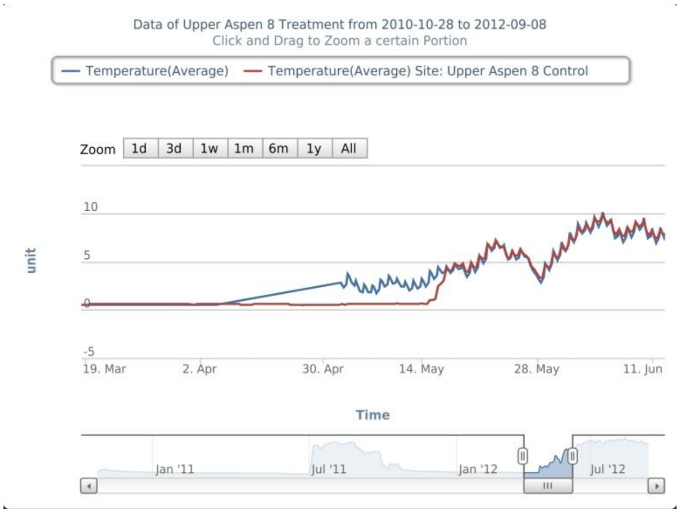


Figure 5. Sample of the data visualization tool in HydroServer Lite, which allows for data assurance as data files are uploaded to the database.

The immediate benefit of having the HSL relational database is that we are transferring the hordes of data that previously resided on our lab computers in multitudes of separate files to a single location where they are linked together to facilitate the types of analyses we need to perform on them. Our experience setting up and using the HSL database has been a catalyst for instituting other data management practices. For example, we found that loading data files into HSL is much easier when the files are systematically named, and we have started to systematically name all of the data we collect from our experiments.

D. *Challenge 2: Short-term data loss*

In addition to collecting sensor data, the Gill Lab also collects field measurements for dozens of different variables. These data are typically collected by Master's and Doctoral students. A particular challenge is the contribution of novice undergraduate researchers who are typically with the lab for one or more seasons. The second data-management challenge we face, short-term data loss, arises from a number of different causes. Data have been lost when the lab computer crashed last summer and when a flash drive containing data files was misplaced. They have also been lost because data were entered into different files and these files were not associated or combined. We also loose data when students leave the lab and have not provided sufficient metadata to make the data they collected useful to their replacements. Solutions to the problem of short-term data loss are encapsulated in best practices associated with data collection and data description.

E. *Best Practices for Collection and Description*

Best practices for data collection require the integration of collection with description so that metadata are created along with the measured data and then stored and transferred together with the data files, so that the data do not become meaningless and the context of their creation is preserved [3]. This integration of collection and description requires extra planning and effort upfront and additional attention throughout the file creation and sharing process.

One of the first steps to integrating collection and description is to enter metadata on standard field-sampling sheets. When data-collection sheets are setup before going out to the field they can contain much of the essential metadata. The sheet may be titled with the measurement being collected, and have spaces to record the names of the people collecting the data, and the date. Site information such as the latitude, longitude, elevation, vegetation class, soil type, and aspect may also be included. It may be important to record information about the weather or the time of day. There should be space to record notes next to each observation, so data collectors can write down any changes or anomalies they notice.

Data collection and description will also improve if the people who collect the data understand how to use the data sheets and appreciate the long-term significance of the data. Spending time on training in the proper collection and recording of data may save time later during the quality control and assurance process. Having data sheets prepared ahead of time for the data that need to be collected can help to ensure that data are not missed in the field.

Datasheets provide a hard copy of the data that can be copied and stored either as a paper copy or scanned and saved as a .pdf or an image file. This preserves all of the metadata and data values together in their original arrangement and context. When transferring the data values to spreadsheets these files should replicate the original data-collection sheets. After the data are entered they should be saved to an external hard drive, an online drive, or a DVD, and stored away from the original computer. One copy of the data should be stored off site or in a fireproof container, so that if any accident or catastrophe occurs and the data are lost from the original computer there is always a reliable backup.

Best practices for file management suggest that the lab establish a data backup policy that includes documentation of data backup [19]. Data stored on electronic media (CD, DVD, or an external hard drive) should be regularly tested and replaced as needed. Electronic media should be stored away from dirt, water, extremes in heat, and light. Files should be saved in open formats (.csv or .txt) rather than proprietary formats (.xls) that require specific software or specific versions of the software [14,20]. The metadata files should describe any software requirements needed to use the data and should indicate which version of the software was used to create the data. Data files should contain one or more header lines that include relevant metadata and also describe the meaning of the contents and the units that apply to each data column.

F. *Our collection, description, and data backup plan*

For short-term data management, including collection, description, and data backup we have a multi-step plan. First, we have included data-collection protocols and data-collection sheets on our lab website,[7] so that students going to the field to collect data will follow the same protocols and collect the same types of data and metadata together.

[7] http://gilllab.byu.edu/

Second, when students bring a data-collection sheet back to the lab and enter the data, or when they create a file in the lab to record measurements, these files are backed up in the "Gill Lab Data in Progress" folder in Google Drive. This folder is shared by the lab and is organized hierarchically by experiment and type of data being collected. Files within the drive have a standard naming protocol, so each file is named by the experiment, data type, and date. Each data file has metadata in the header that provides the metadata from the collection sheet and describes the meaning of the values in each column of data. A scanned copy of the original datasheet is saved with the file and given the same name. Any scripts that were used to clean, process, or analyze the data are also stored with the data in this file. The hard copies of the original data sheets are either kept by the graduate student, if part of their thesis or dissertation work, or kept and filed in the lab at the professor's discretion. The Google Drive that is shared online is synced regularly with the PCs of the users, so there are multiple dispersed versions of the data that are automatically backed up and that replace each other as data are added or previous versions of the files are updated.

Finally, when the data sets are complete, the student(s) who created them completes a metadata text file that describes why the data were made, how they were used, who created them, the methods that were used, and other information according to pertinent metadata standards. A metadata editor may be used to help complete this file. The professor or graduate student overseeing the project reviews and edits the metadata files and indicates that they have done so with a date. Then, the files are stored in the "Gill Lab Final Data" folder on the Google Drive. When appropriate these final data files are shared on the lab website or deposited in a data archive.

G. Challenge 3: Image archiving

We have recently begun two projects using high-resolution photogrammetry that generate thousands of images that are used to build 3-D models of soil surfaces and plant canopies. In addition, on nearly every trip we take to the field to collect data we also take pictures that capture landscape and individual conditions. These images can provide valuable information for interpreting the data we collect and in many cases the images themselves are the data. For some experiments, we use repeat photography to measure fine-scale changes in leaf area index or canopy coverage. Similar to the flood of sensor data, digital images quickly become numerous and uninterpretable. Additionally, we often need to compare information contained in several different images, and to do this we convert the images to GIS data layers or use image management software such as ImageJ [21]. This means that multiple files and file types may be needed to interpret a collection of images. Images naturally contain some metadata, such as the date of creation and specific camera settings, but there are other metadata that we want to keep with the images that are not automatic, so we need a solution that associates metadata with the images and facilitates the search and retrieval of related files.

H. Best Practices for Images and Multimedia

Many options exist for archiving and sharing images. Web-based image galleries such as Facebook, YouTube, Picassa, Vimeo, or Flickr are available at low or no cost and allow for tagging images with metadata and organizing files. These may provide valuable avenues for sharing images and protecting original copies against data loss [22]. Commercial image-management systems have been designed for professional photographers. Examples include Adobe Photoshop Lightroom and Photo Mechanic [23,24]. These may be useful for managing images within the lab, but may be problematic for sharing metadata if the receiving party does not also own the software. Image management systems such as Morphbank[8] have been created as repositories for certain types of images, in this case for specimens from museums and herbaria. These systems require specific metadata and may not be useful outside of the collections which they were created. Still, they provide examples of working solutions to multimedia archiving.

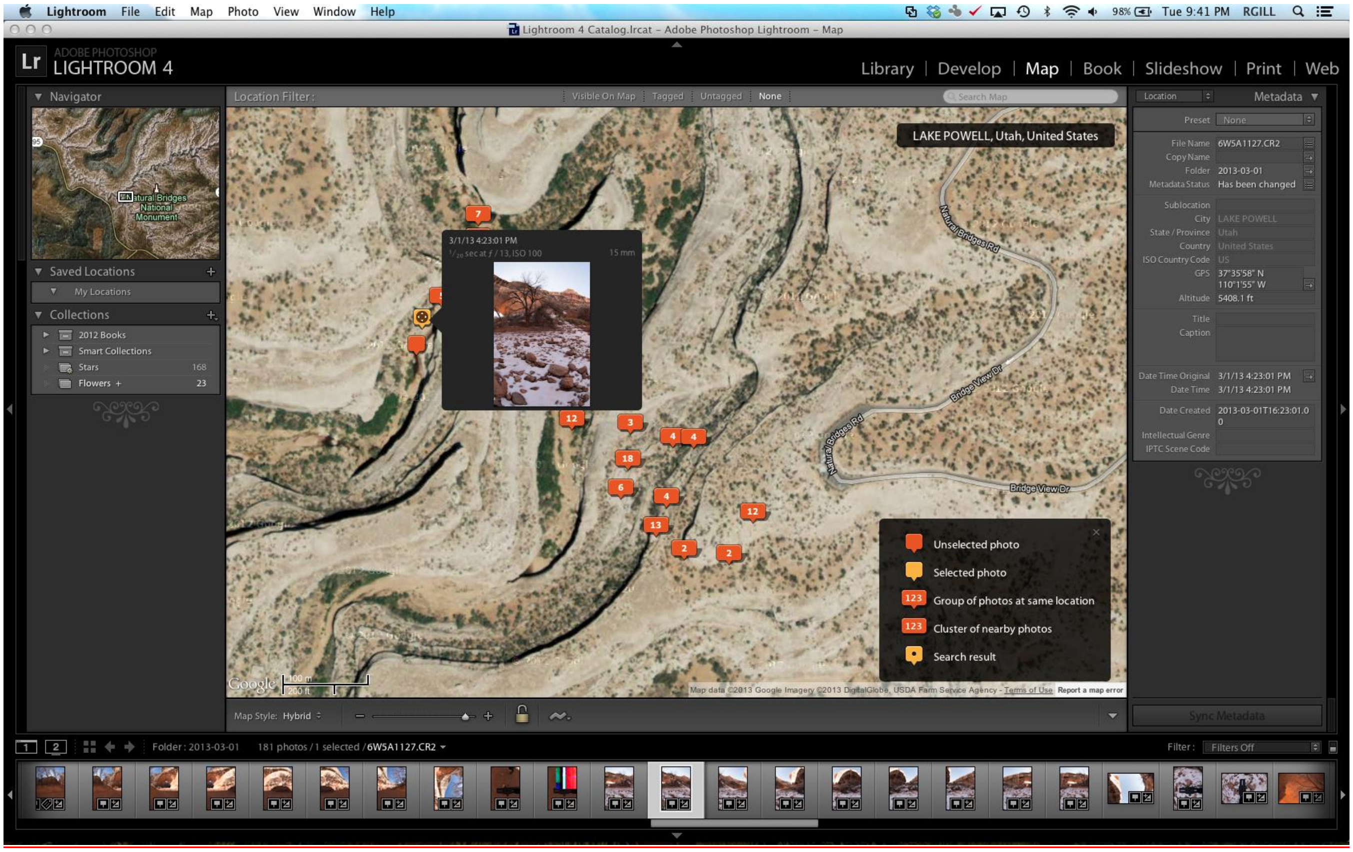
The best practices for manageing and preserving still images, moving images, and sound recordings are to save the original copies of the files in the best quality possible, and to store copies of the original files on a separate drive other than the one from which images are routenly used for analyses [22]. This practice maximized the information stored in the files and protects the original files from being lost.

I. Our image management plan

In the Gill Lab, we are adopting a number of solutions to image management. For simple image sharing we have used Dropbox, Facebook, and email. This makes image files available from any computer and eliminates the loss of images from failed drives. But, the limitation of Dropbox is that it has a storage size limit, which requires the purchase of additional storage space. In this way we can associate some additional types of metadata with our images. But, this solution is mainly used for a small number of images and typically those that are used in presentations and not the ones that are actually data.

As with other data, image management begins prior to data collection. First, we have established the standard of collecting all research-related images in RAW format to maximize flexibility in analysis and the maintenance of as much spectral information as possible. Images are captured with a camera identifier and date prefix with sequential numbering to ensure that each image has a unique file name. When images are downloaded, they are individually evaluated and metadata generated. We have used Adobe Lightroom as the primary image management program (Figure 6). Image metadata include annotating each image with standardized keywords for experiment and location information. Additional keywords are added as needed. With Lightroom, we can add these keywords to entire batches or collections of files, simplifying annotation of photos. For most of our images, we automatically georeference the image using a GPS receiver on the camera hotshoe. These data are automatically included in the image

[8] http://www.morphbank.net/


Lightroom File Edit Map Photo View Window Help
98% Tue 9:41 PM RGILL
Lightroom 4 Catalog.lrcat - Adobe Photoshop Lightroom - Map
Lr ADOBE PHOTOSHOP LIGHTROOM 4
Library | Develop | Map | Book | Slideshow | Print | Web
Navigator
Natural Bridges National Monument
Saved Locations
My Locations
Collections
2012 Books
Smart Collections
Stars 168
Flowers + 23
Location Filter: Visible On Map Tagged Untagged None
Search Map
LAKE POWELL, Utah, United States
3/1/13 4:23:01 PM
1/20 sec at f / 13, ISO 100 15 mm
Natural Bridges Rd
Bridge View Dr
Unselected photo
Selected photo
Group of photos at same location
Cluster of nearby photos
Search result
Google 100 m 200 ft
Map data ©2013 Google Imagery ©2013 DigitalGlobe, USDA Farm Service Agency - Terms of Use Report a map error
Map Style: Hybrid
Location Metadata
Sync Metadata
Filter: Filters Off

Folder: 2013-03-01 181 photos / 1 selected / 6W5A1127.CR2

Preset: None
File Name: 6W5A1127.CR2
Copy Name:
Folder: 2013-03-01
Metadata Status: Has been changed
Sublocation:
City: LAKE POWELL
State / Province: Utah
Country: United States
ISO Country Code: US
GPS: 37°35'58" N 110°1'55" W
Altitude: 5408.1 ft
Title:
Caption:
Date Time Original: 3/1/13 4:23:01 PM
Date Time: 3/1/13 4:23:01 PM
Date Created: 2013-03-01T16:23:01.00
Intellectual Genre:
IPTC Scene Code:

Figure 6. Screen shot of Adobe Photoshop Lightroom showing geo-referencing and metadata tools, photos by R.A. Gill taken at Natural Bridges National Monument, San Juan County, Utah 2013.

metadata file. While the keywords are added in in Lightroom, they are associated with the file and can be read in any RAW compatible image viewer.

After images are indexed they are backed up to a local external drive and stored on DropBox. Annually they are burned to a DVD for long-term storage and deleted from DropBox due to space limitations. We always maintain two copies of each RAW image. Many of our applications use images that have been converted and exported as a jpg. Processed files are maintained by individual investigators based on their application needs. However, the RAW files remain untransformed in storage.

IV. Conclusions

Data availability is changing the way we do science. Reseach and scientific dicovery increasingly involve the discovery and intergration of widely available digital stores of data, and the traditional data lifecycle is not suffient to support science and eduation through the 21st century. In addition to the digital stores of data from national and international projects, individual labs are also capturing new types of data in ever increasing volumes. These independent research labs need collaborations and support from specialists in fields of data informatics, who are provided open system solutions that can be learned and implemented quickly and customized for the individual needs of the lab and that also comply and enforce compliance with the data management standards specific to the research community.

In the Gill Ecosystem and Global Change Ecology Lab, we have benefited from collaboration with specialists in hydroinfromatics who provided the HydroServer Lite database as an open system data-base platform for storing, querying, and sharing data from environmental sensors. We are also implementing other data management solutions that meet the varied requirements for short-term data preservation and intergration and that will faciliate long-term storage and the future discovery of the data. We hope that through these data management practices our lab will remain competitive and relavent in an increasingly tight funding environment. We also hope that our data will support larger scientific endeavors and future discroveries.

Acknowledgment

L.G.C. thanks Dan Ames who taught the Hydroinformatics class that introduced him to data management and for helping to set up the HydroServer Lite website and to thank Jiri Kadlec for help troubleshooting the installation and registry of the HydroServer Lite database. We would also like to thank C. Riley Nelson for sharing data management ideas especially about image archiving.

References

[1] T. Hey, S. Tansley, and K. Tolle, Eds., *The fourth paragigm: data-intensive scientific discover*," Redmond, WA: Microsoft Research, 2009.

[2] M. Nielson, *Reinventing discovery: the new era of networked science*, Princeton: University of Princeton Press, 2012.

[3] C.A. Strasser, S.E. Hampton, "The fractured lab notebook: undergraduates and ecological data management training in the United States," *Ecosphere* vol. 3, article 116, December 2012.

[4] C. Strasser R. Cook, W. Michener, A. Budden, R. Koskela, "DataONE: promoting data stewardship through best practices. In *Proceedings of the Environmental Information Management Conference 2011*, M.B. Jones, C. Gries, Eds. University of California, 2011, pp. 126-131, doi:10.5060/D2NC5Z4X.

[5] C. Strasser, R. Cook, W. Michener, A. Budden, "Primer on data management: what you always wanted to know," Accessed April 2013. http://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf.

[6] National Science Foundation, "Disemination and Sharing of Research Results," Accessed April 2013, http://www.nsf.gov/bfa/dias/policy/dmp.jsp.

[7] National Science Board, "Long-lived digital data collections: enabling research and education in the 21st century," Technical Report NSB-05-40, National Science Foundation, September 2005, http://www.nsf.gov/pubs/2005/nsb0540/.

[8] T. Hey, S. Tansley, and K. Tolle, Eds., "Jim Gray on eScience: a transformed scientific method," in *The fourth paragigm: data- intensive scientific discover*," T. Hey, S. Tansley, and K. Tolle, Eds. Redmond, WA: Microsoft Research, 2009, pp. xix–xxxii.

[9] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, S. Mock, "Kepler: an extensible system for design and execution of scientific workflows," 16th International Conference on Scientific and Statistical Database Management (2004) IEEE publication number P2146

[10] ArcGIS Desktop, Redlands, CA: Environmental Systems Research Institute, 2013.

[11] Python. Beaverton, OR: Python Software Foundation, 2013.

[12] R Core Team, R: A Language and Environment For Statistical Computing, Vienna, Austria, 2013.

[13] J. H. Porter, P.C. Hanson, C. Lin, "Staying afloat in the sensor data deluge," *Trends in Ecology and Evolution* vol. 27, pp. 121-129, February 2012.

[14] DataONE, "Preserve information: keep your raw data raw," DataONE Best Practices, Accessed April 2013, http://www.dataone.org/best-practices/preserve-information-keep-your-raw-data-raw.

[15] DataONE, "Identify values that are estimated," DataONE Best Practices, Accessed April 2013, http://www.dataone.org/best-practices/identify-values-are-estimated.

[16] DataONE, "Confirm a match between data and their description in metadata," DataONE Best Practices, Accessed April 2013, http://www.dataone.org/best-practices/confirm-match-between-data-and-their-description-metadata.

[17] J. Kadlec, D.P. Ames, "Development of a lightweight HydroServer and hydrologic data hosting website," in Proceedings of the AWRA Spring Specialty Conference on GIS and Water Resources, New Orleans, LA, 2010, http://his.cuahsi.org/documents/JiriKadlec_a80004b9_7940.pdf.

[18] D.P. Ames, J.S. Horsburgh, Y Cao, J. Kadlec, T. Whiteaker, and D. Valentine, 2012. HydroDesktop: Web Services-Based Software for Hydrologic Data Discovery, Download, Visualization, and Analysis. Environmental Modelling & Software. Vol 37, pp 146-156. http://dx.doi.org/10.1016/j.envsoft.2012.03.013.

[19] DataONE, "Create and document a data backup policy," DataONE Best Practices, Accessed April 2013, http://www.dataone.org/best-practices/create-and-document-data-backup-policy.

[20] DataONE, "Document and store data using stable file formats" DataONE best practices," Accessed April 2013, http://www.dataone.org/best-practices/document-and-store-data-using-stable-file-formats.

[21] Wayne Rasband, ImageJ, Bethesda, MD: National Institute of Mental Health, 2013.

[22] DataOne, "Plan for effective multimedia management," DataOne Best Practices, Accessed April 2013, http://www.dataone.org/best-practices/plan-effective-multimedia-management.

[23] Adobe Photoshop Lightroom, San Francisco, CA: Adobe Systems Inc., 2013.

[24] Photo Mechanic, Portland, OR: Camera Bits Inc., 2012.