Proposed venue: **Library Hi Tech News**

Title: **Interactive data entry and validation tool: A collaboration between librarians and researchers**

Author:

Abstract
Purpose – To share a case study process of collaboration with faculty, highlight some design considerations when creating data entry forms, and describe some considerations when planning for data management.
Design/methodology/approach – A case study is presented with goals, events and design considerations.
Research limitations/implications – Each partnership is different with different data management needs. Design considerations for a data entry form depend on the specific data being collected.
Practical implications – Principles can be applied to other libraries that are developing data management collaborations with faculty, or for designing of web/paper data entry forms.
Social implications – Good data collection and storage for future access promotes reuse of data for additional research.
Originality/value – Typically forms are not interactive beyond simple presentation of alternate questions. This form design method builds on user inputs to create dynamic, color coded and textual guidance. Some of the process of collaborating with faculty partners is shared.

> "IT professionals…. need to take the view that data is a precious thing and will last longer than the systems themselves." —Tim Berners-Lee (Runciman, 2006)

Introduction

Instead of being just guardians of the books, librarians with their digital information organization skills are also active, collaborative partners in the faculty research process. The Center for Digital Scholarship in the Miami University Libraries partners with faculty researchers to provide data management support. This article describes a collaborative design process of creating system to collect and store data, focusing on data collection and validation, and highlights some principles of data entry design.

History

A faculty research group we had previously supported asked us for a new letter of expanded support to accompany their data management plan in a National Science Foundation (NSF) grant application. The faculty group's university computer and technology support team had been reorganized in the previous year, creating an increased need for data validation and storage that the library could provide. The library committed to provide support for all phases of

the project from data collection, metadata design, and active storage for primary researchers through post-project storage and public availability.

A data management plan is a carefully thought out series of policies and procedures for every step of data's life. The NSF requires collected data to be easily available for additional research. For no extra grant dollars, additional research analysis can be conducted, adding to humanity's stock of knowledge.

> Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. http://www.nsf.gov/bfa/dias/policy/dmp.jsp("US NSF - About," n.d.)

To do this, the researcher needs to consider several things:

- How are the data collected
- Creation and recording of the associated record level metadata to facilitate searching within the data set
- The data responsibilities of each member of the research team
- How the data will be stored and accessed during the project
- How the data will be consolidated for public research
- Where the data will be housed and made available after the project
- Associated collection level metadata so that the whole data store can be found

Because librarians are well equipped with the conceptual foundations for making information accessible, they are not only natural partners with researchers but can also be valued members of a research team.

Process

Communication with the faculty research team to discover the team's needs was followed by a letter of support from the libraries for the grant application. We then developed both paper and electronic data collection tools, a database and modified the workflow for their data.

Project Description

Several decades of zooplankton lake data have been collected. Data includes weather, temperature and other physical parameters at varying depths, and zooplankton specimens for later identification and counting. Information was embedded in written comments as well as spread through a variety of spreadsheets and data files.

The data had been initially recorded on paper forms that were then transcribed into spreadsheets. Transcription was done by different personnel several months after the data had been collected. Researchers expressed concerns to the library about potential data entry

errors. They had been working on a data validation workflow that could identify possible errors in the spreadsheets. For example, if the depth for a measurement was recorded as 25 meters but the lake was only 15 meters deep, the program would indicate that there was an error. Further human analysis could sometimes guess a number, but otherwise the correct value was permanently lost.

Raw data spreadsheets were processed with R (a statistics computer language) to produce summary spreadsheets. Extracting information from the summary data required a series of steps. Spreadsheets would be filtered and sorted; an R program would be run to consolidate various readings followed by further calculation and sorting to produce final values. Repeating the same data extraction from a different time period or for a different variable required manual repetition of each of the steps.

One of the goals of our partnership became the merging of the various spreadsheets and data files into a few consistent database tables. When the data are combined into a database, it is possible to create a Structured Query Language (SQL, previously known as SEQUEL) query to extract information. Each time the same type of information is needed, the same SQL query (or set of queries) can easily be rerun. Since SQL is human readable text, changes in parameters to extract different information are simple to implement.

Designing data collection forms to prevent errors is an art. An entry form is a conversation between the researcher and the person filling the form (Jarrett and Gaffney, 2009). Form filling errors can be omission, commission, mistakes, and transcriptional. Requiring an entry can prevent many omission errors. Controlled vocabulary and guiding instructions can limit commission errors. Validation routines can reduce some types of mistakes. Avoiding handwriting by directly typing the answers removes transcription errors. While some of these techniques are available for paper forms, active software can be stronger in guiding the data collector and reducing errors.

By checking the data for validity when initially recorded, the person collecting the data can provide the correct value immediately. Long term human memory limitations, handwriting and many typographical errors could be eliminated (Miller, 1956).

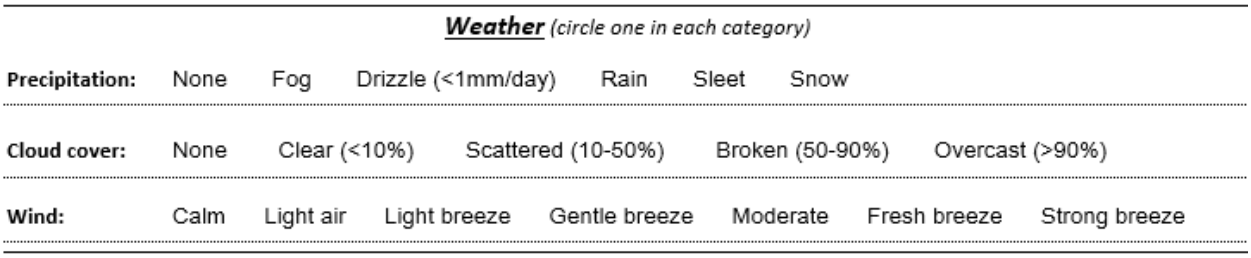
While a software application was desirable, it requires time to develop. There was an immediate need to collect data, so a two phase approach was taken. The first phase was to improve the paper form for that summer's data collection with a software based approach being developed for the following year.

Form design

The generally good two-sided paper data collection form had been developed over several years. Entry boxes were in the same order as the data collection procedures. To improve it, several changes were made.

Informative Fonts:
Each font indicated a different type of information. Instructions to the data collector, headings, and check options were respectively in their own font. This reinforcement helps the data collector quickly understand the purpose of individual lines in the form.

Circle-the-answer and Check boxes instead of Fill-in-the-Blanks:
When possible, a list of acceptable answers was provided to the data collector. That controlled vocabulary guided the data collector and prevented handwriting transcription errors. Instructions were used to remind the data collector how to interact with the form. Responses were selected by circling words rather than by handwriting text.



Weather *(circle one in each category)*

Precipitation:	None	Fog	Drizzle (<1mm/day)	Rain	Sleet	Snow	
Cloud cover:	None	Clear (<10%)	Scattered (10-50%)	Broken (50-90%)	Overcast (>90%)		
Wind:	Calm	Light air	Light breeze	Gentle breeze	Moderate	Fresh breeze	Strong breeze

Figure 1 Guidance within each question

Spacing and horizontal rules:
Each task in data collection was separated visually on the page to make it easier for the data collector to find the correct field to fill. Data entry arrangement matched the collection procedure sequence.

Data fields instead of comments when possible:
Instead of having text about the weather condition in a comment, specific fields were used for each characteristic to be collected such as rain, sun and wind. Guidance for the meaning of each of the allowed characteristics was included either in the answer, or in a separate section on the same sheet of paper. Standard terms were used so that the weather could be compared to conditions recorded over the centuries. The controlled vocabulary also allows the weather data to become part of the search context when using SQL. We retained the ability to record comments elsewhere in the form for non-standard issues that may arise.

Identifying information on each page:
Since sheets may be photocopied or get separated, the top of each page includes key identification elements. In this case the key values are the location and Date.

Highlight and guide unfamiliar items:
The date format was changed to a more sortable and international format of year followed by month then day. To remind collectors, the format was highlighted in yellow with light gray guiding characters in each field.

Figure 2 Identification section at the top of each page

Provide a comments section / Provide additional guidance and instructions:
A section for data collector comments was specified. It also serves as a section for guidance and instruction by having definitions of wind speeds in faint text that the data collector can simply write over. This removes the need to have a separate instruction sheet that can easily be misplaced.

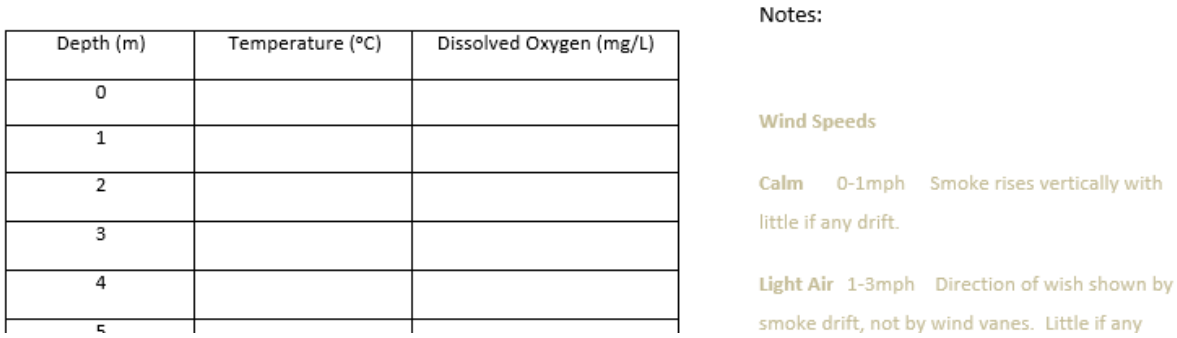


Figure 3 Notes section

Software interface design
The second phase of data collection support included the development of software tools for data collectors. Both Android and Windows versions were created. Since the field teams were already using PC laptops for instrument monitoring, they preferred the Windows version. The software was designed to respond actively to the user's input, not just passively recording their entries, thus enabling advanced features.

- Immediate data validation can prevent data entry errors resulting from incorrect values, typos, and handwriting interpretation. It can also provide in-the-field training to the data collector, improving future data entry. The time and date of data collection can be automatically recorded.
- Data validations can be context sensitive. For example, after a field such as Lake Name is entered, other fields can be pre-filled (such as geographic region name) or used in validation (maximum depth of that lake must be greater than the sampling depth).
- Since the data are in electronic form and the software is aware of appropriate metadata for each entry, the data can be automatically uploaded to a central server. This means

the primary researcher back in the lab can begin analysis immediately rather than waiting months for data transcription to be completed.

To develop the software interface, the same layout concepts from the paper form design were used. The goal is to guide the data collector, making it simpler to accurately record data. In addition to spacing, fonts, and controlled vocabulary; tabs and colors were added. Each major section (corresponding to each major data collecting activity) was color coded and put under an ordered tab. Multi-state toggle buttons and drop-down lists replaced check boxes.

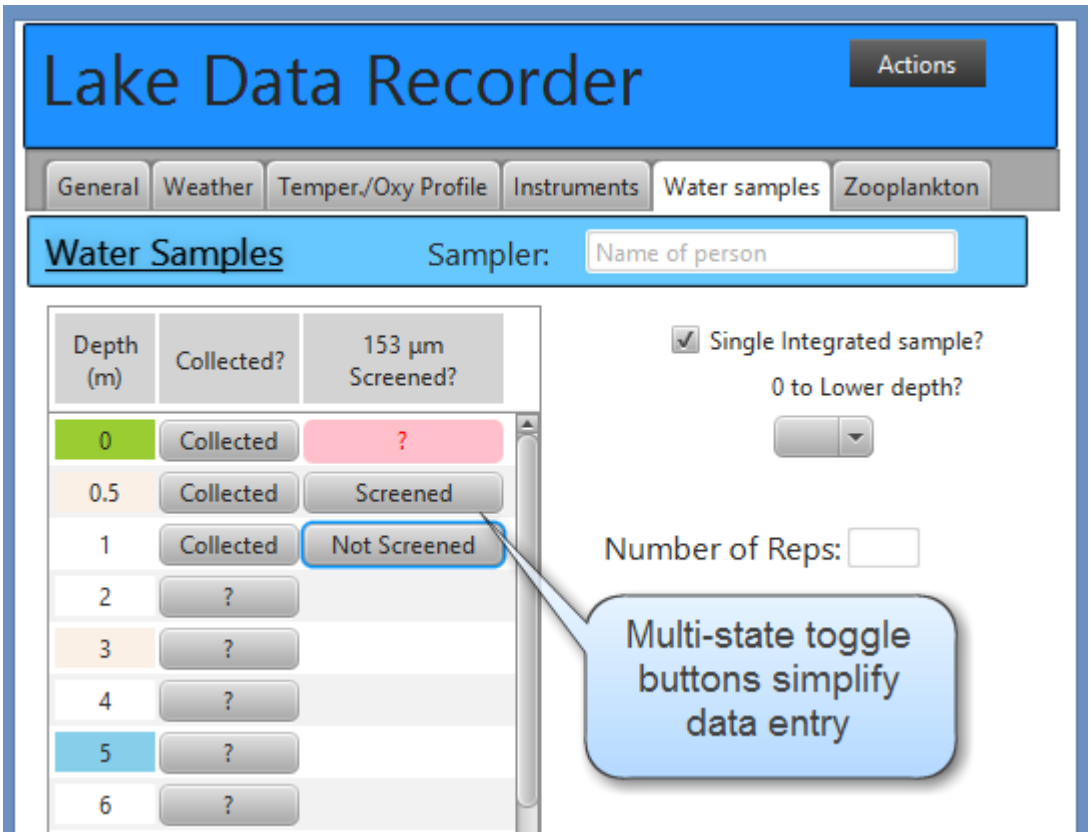


Figure 4

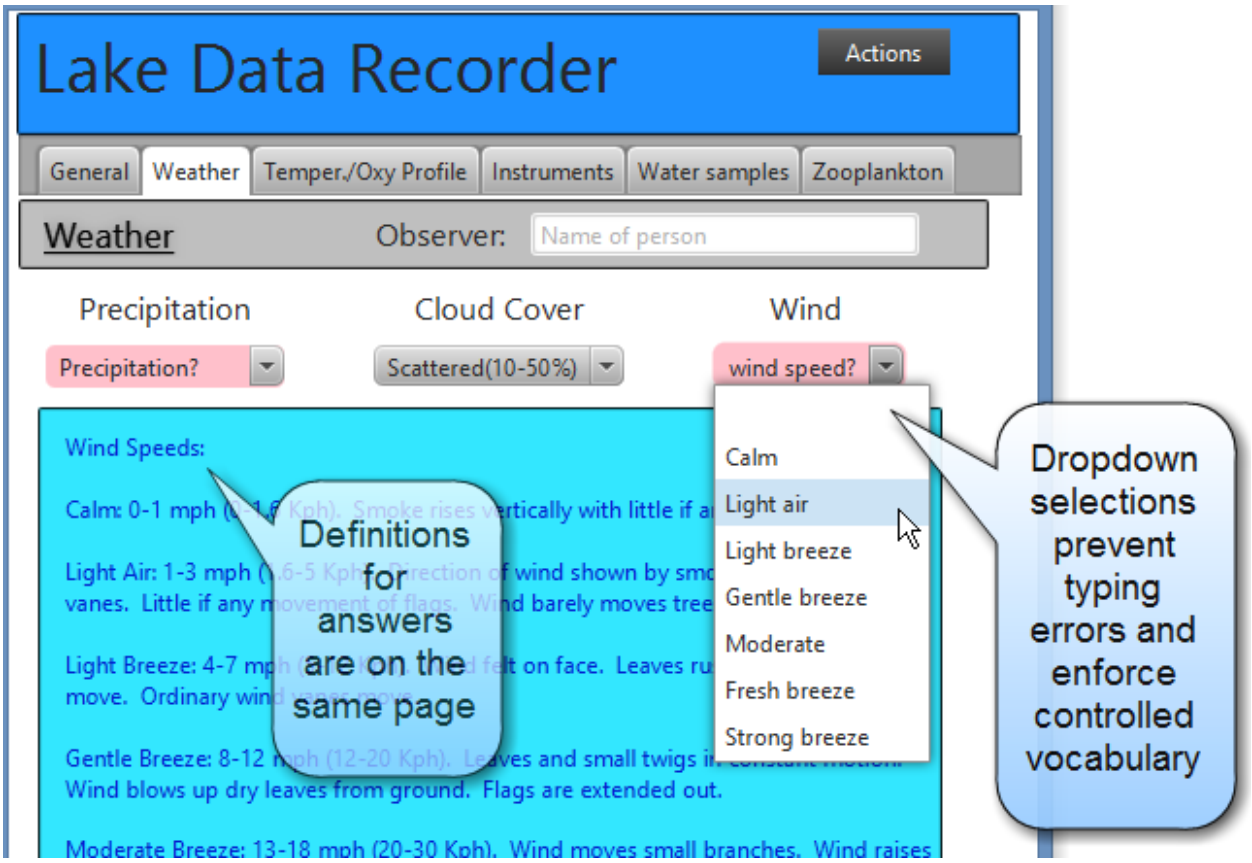

Lake Data Recorder
Actions
General
Weather
Temper./Oxy Profile
Instruments
Water samples
Zooplankton
Weather
Observer:
Name of person
Precipitation
Cloud Cover
Wind
Precipitation?
Scattered(10-50%)
wind speed?
Wind Speeds:
Calm
Light air
Light breeze
Gentle breeze
Moderate
Fresh breeze
Strong breeze
Definitions for answers are on the same page
Dropdown selections prevent typing errors and enforce controlled vocabulary

Figure 5

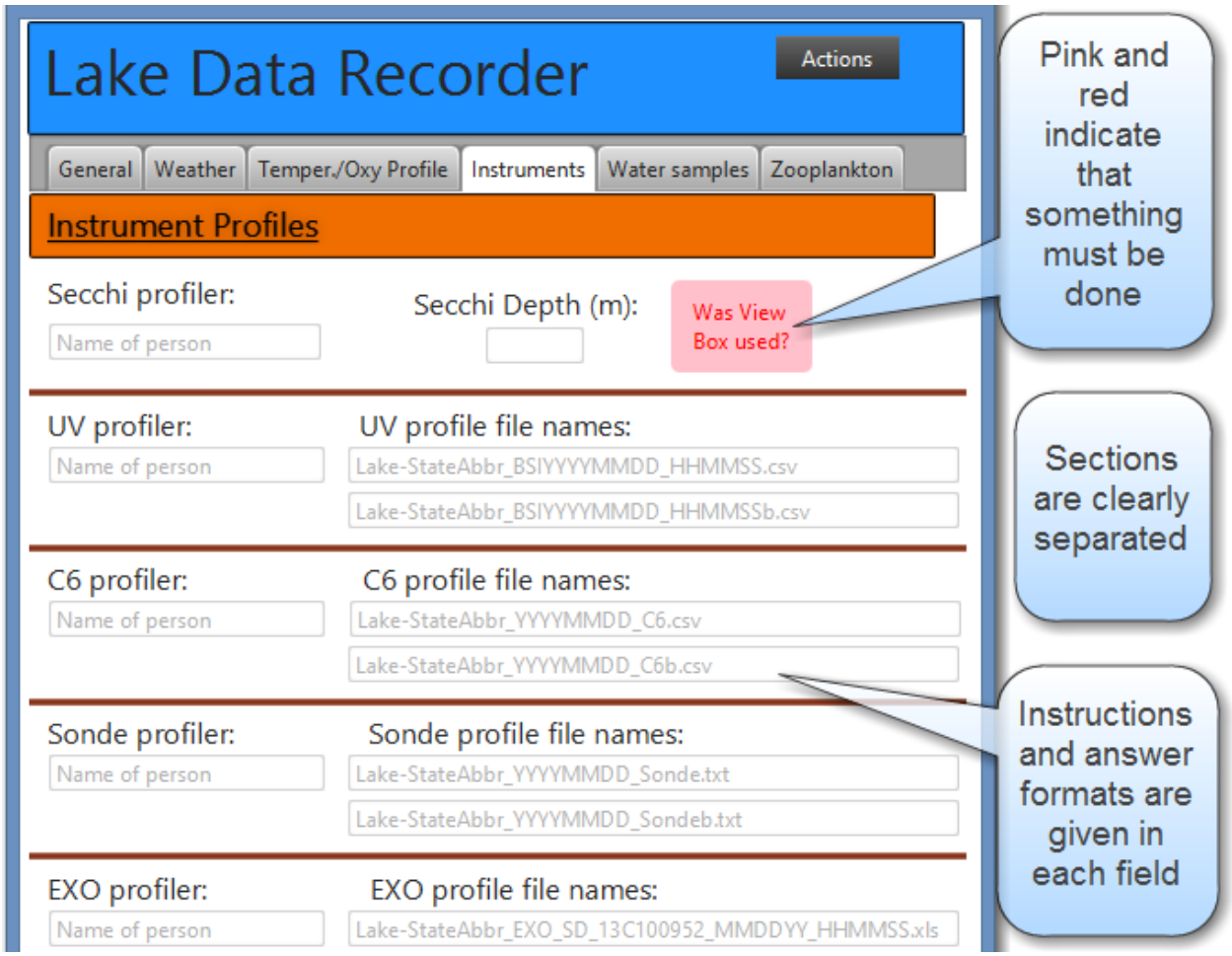


Figure 6

Data from some fields were used to automatically fill out other fields. In this example, the lake name triggered filling the Country, State, Time and Date fields.

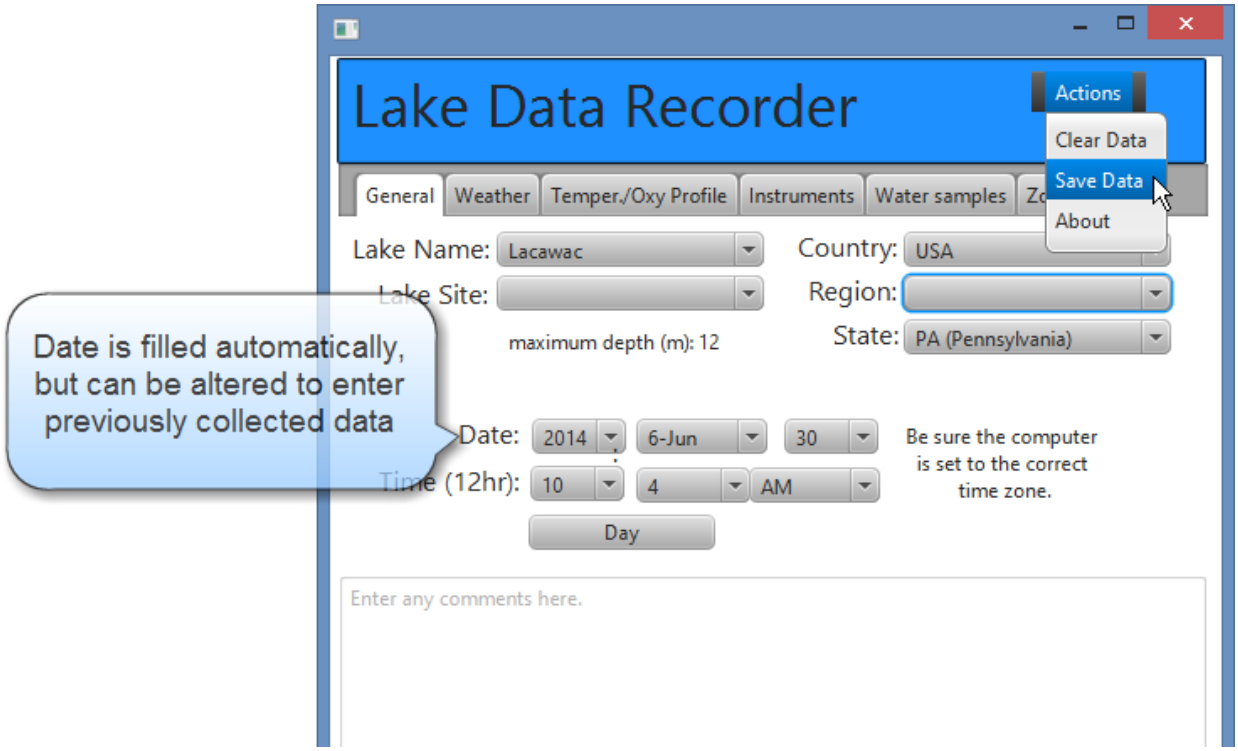


Figure 7

A value entered can also cause a portion of the form to appear or disappear as needed.

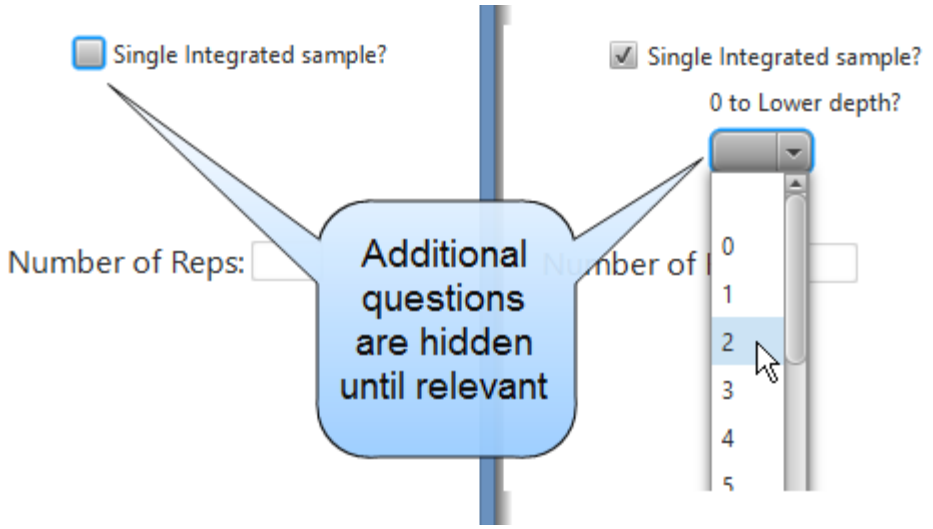


Figure 8

Context sensitive active messages and instructions can help guide the data collector. When errors were detected or entry was required, the entry field became pink.

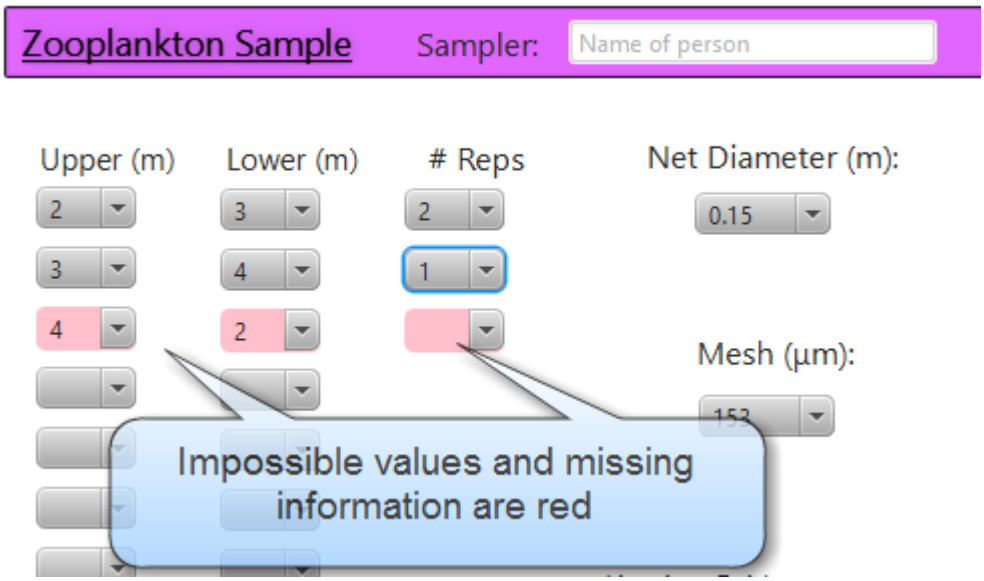


Figure 9

Entered data could be used to give multi-colored guidance to the data collector. For example, the temperature of a lake varies with depth. The rate of temperature change indicates different underwater layers. Previously, data collectors would look at the numeric temperature values entered to identify each layer. With active software, the temperatures were transformed into a color coded chart that could be used to identify layer boundaries.

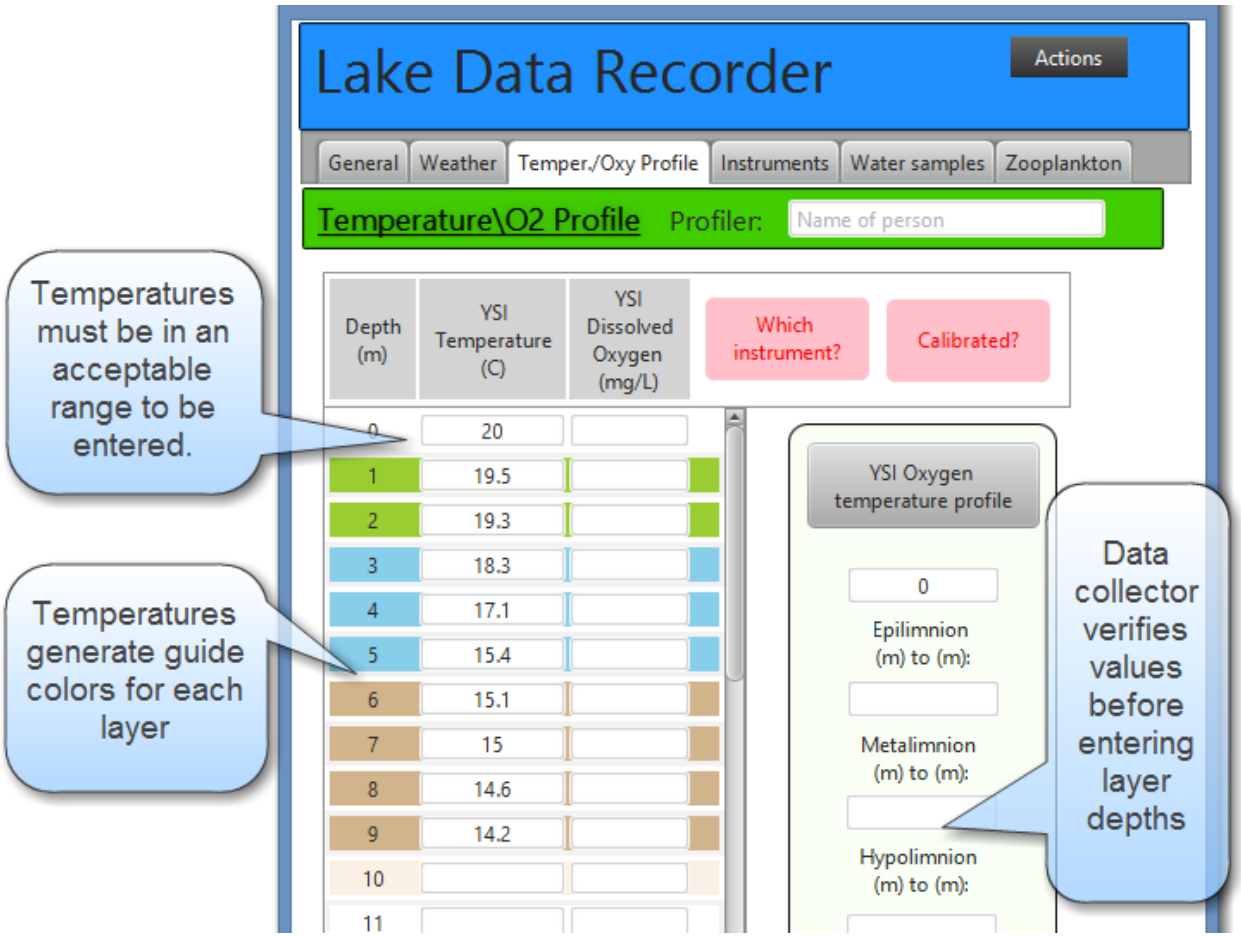


Figure 10

Data storage and access

Data that is collected is stored locally on the field laptop. Each day of collection, a local backup copy is automatically made so that the field researchers can examine and copy it to field backup devices as desired. The software also queries the main server, waiting for a valid connection. When found, the data is uploaded and verified. The local copy of the data is then marked to indicate it has been successfully uploaded.

Data collection is only one step in the process of data management. The metadata for this project and the associated database structure were designed alongside the development of data collection software. The data is stored in a jointly accessible location so that investigators can search and extract data as needed using standard SQL queries.

Once the data is stored, access is through two related primary tables. One table captures site-specific information such as date, location, weather, and other data related to the collection activity. The other related table lists information about conditions at each depth range including dissolved oxygen, illumination, temperature, and identification of samples taken. An additional

table holds parameters used by the software such as validation rules and maximum lake depths. Raw data files produced by data collecting instruments are converted from CSV (Comma Separated Values) files to data tables that can be linked to the primary site table. A supplementary file includes descriptions of each metadata definition known as the "codebook" so that both the current investigators as well as future researchers can understand what each data field means. Standard SQL tools and syntax can be used to query the data and provide some preprocessing.

To help other researchers access the data, the Ecological Metadata Language (EML) will be used. EML is an eXtensible Markup Language (XML) based metadata language used to describe an overview of the data contained in the repository. When combined with the globally developed Morpho data management software and data network servers such as the Knowledge Network for Biocomplexity (KNB), researchers from across the world can access find and then access the data collected by this project.

Summary

Collaborations can be beneficial for both the library and faculty. Form design for data entry should focus on the needs and perceptions of the data collector. Data collection is a conversation that strives to reduce errors and simplify meticulous data collection. Metadata and long term data storage with public access should be addressed from the start. Librarians are well suited to provide this support and to collaborate with faculty about data management needs and practices.

References

Jarrett, C., Gaffney, G., 2009. Forms that Work: Designing Web Forms for Usability. Morgan Kaufmann.

Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63, 81–97.

Runciman, B., 2006. Isn't it semantic? ITNOW 48, 18–21.

US NSF - About [WWW Document], n.d. URL http://www.nsf.gov/bfa/dias/policy/dmp.jsp (accessed 6.27.14).

About the Author