final report

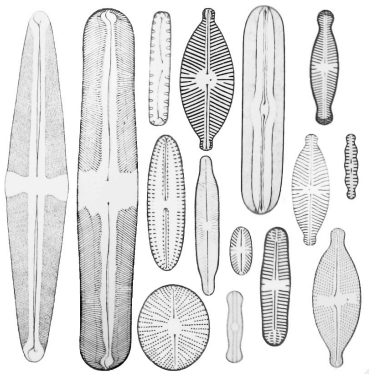
USGS
RM-CESU
CU-BOULDER

TAXONOMIC CONSISTENCY IN BIOLOGICAL ASSESSMENT:
Developing a web-based identification guide & ecological resource for diatoms of western North America

David Lubinski
Investigator

INSTAAR
Univ. of Colorado at Boulder
Campus Box 450
Boulder, CO 80309

303-735-6619
David.Lubinski@colorado.edu



Award Number:	04121HS011
Type of Project:	Technical Assistance
Project Discipline:	Natural
Funding Agency:	USGS
Other Partners:	US EPA
Effective Dates:	4/11/07 to 2/28/09
Funding Amount:	$63,899

October
2009

A project funded in part by the U.S. Geologic Survey through the Rocky Mountains Cooperative Ecosystem Studies Unit to the Univ. of Colorado at Boulder

The team

Sarah Spaulding

project leader (USGS)

David Lubinski

website leader (INSTAAR)

Marina Potapova

diatom specialist (ANSP)

Julie Kinsey

scientist (EPA)

Tisza Bell

grad student (INSTAAR)

Keywords

web-based guide
diatoms
western North America
identification
website, web, site
taxonomy
taxa
genus, species

OVERVIEW

The problem

Diatom species composition and abundance reflect the biotic condition of freshwater streams and lakes. Together with aquatic invertebrates and fish, diatoms are an indispensable component of state and federal water quality monitoring and assessment programs. The diatom data created by many of these programs - valuable as they are - can only be used within a particular state or dataset; they are not comparable to data from other states and federal agencies because of inconsistent taxonomy across different labs. This inconsistency is due in part to a lack of systematic resources for identification of western North American diatoms, which routinely forces diatom analysts to use European taxonomic keys. Taxonomic inconsistency has also hindered the determination of ecological tolerances to stressors and the recognition of indicator species.

The solution

With funds from the Rocky Mountains Cooperative Ecosystem Studies Unit (RM-CESU), David Lubinski (INSTAAR - Institute of Arctic and Alpine Research, University of Colorado at Boulder) was able to work closely with Sarah Spaulding (USGS) on her EPA-funded project to design a web-based identification and ecological guide for diatoms of western North America. This guide builds on previous work and success of the NSF-funded web resource Antarctic Freshwater Diatoms (http://huey.colorado.edu/diatoms) created by Spaulding, Lubinski, and others in 2005. The new guide takes advantage of the first large diatom dataset for western North America, created by EPA's EMAP - Environmental Monitoring and Assessment Program. The guide integrates diatom taxonomy, nomenclature, reference images, maps, sample information, species count data, literature references, and ecological analysis. The guide promotes taxonomic consistency between state and federal programs, including the USGS NAWQA - National Water Quality Assessment. In particular, it allows agencies to implement EPA recommendations to accurately assess diatom assemblages in streams and wadeable rivers; thereby assisting their compliance with the Clean Water Act. The guide built for this project is a pilot, populated with 25 key species. Hundreds of additional species will be added in the coming year through additional funding from USGS and EPA (see "future developments" on the last page of this report).

Lubinski's role

Lubinski led the design and programming of the website, including the database that drives the site as well as the site interface, documentation, data flow and data entry, and web security. The RM-CESU agreement enabled Lubinski to hire Bell, a CU-Boulder graduate student specializing in diatom research, to assist with collection and assembly of taxonomic information and creation of imagery. The report you are reading now focuses on Lubinski's role, which was solely funded by this RM-CESU award, including computer software and hardware costs. An additional report on the broader project will be submitted to EPA by Spaulding by 15 February 2010.

OUR PROCESS

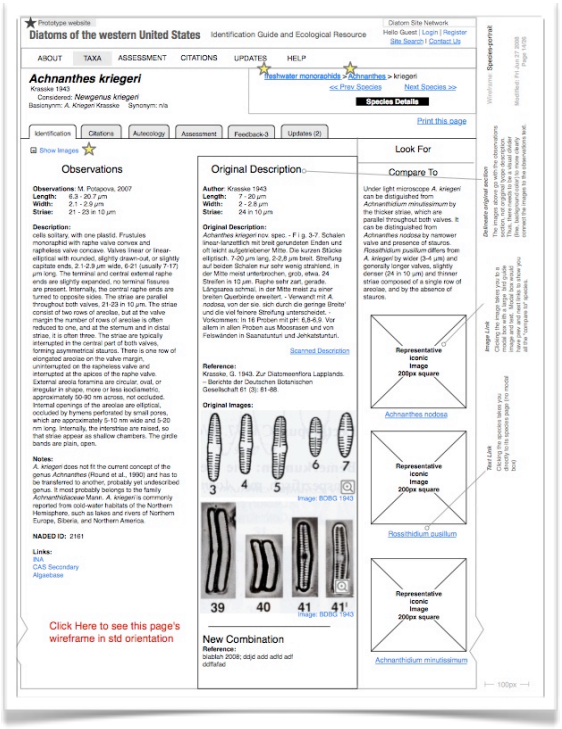
Lubinski led the team through a ten-step design process, each step of which is discussed in detail below, including milestones and outcomes. Although the steps are discussed linearly, the actual process was sometimes non-sequential with intersecting and iterative tasks. Please note that first five steps were intensely collaborative between team members and took much longer than the last five.

1. Research, brainstorming & strategizing

To start the design process, Spaulding, Lubinski, Kinsey, and Bell held a series of intensive brainstorming and planning sessions. We assessed the potential content for the site, investigated analogous websites (especially database-driven identification guides), and refined the purpose and goals of the site. Emphasis was placed on the user experience, while also maintaining biologically accurate content. Audiences were divided into non-expert users (state and federal agency biologists, consultants, and the public) and expert users (diatom taxonomists and ecologists).

These early discussions helped us articulate our approach to the new website, providing a firm basis for making subsequent design decisions. Here is a summary of our approach:

The main goal of our work is to provide the means for federal, state, university programs to reach taxonomic consistency. We are taking an approach of presenting images of diatoms, as they are found in rivers and lakes of the US. These specimens may not (in many cases, clearly do not) fit the Eurocentric naming scheme that, up until now, diatomists have had little choice but to apply. Our approach is pragmatic, in that we present diatom images and experts' interpretation of the name to be used in biomonitoring programs. Because we accept taxa that have not yet been formally described and refrain from forcing the use of existing names, we are uncovering the unique flora of North America. To help keep the user interface easy to use, we will not employ complex taxonomic keys. In the inevitable tradeoff between interface efficiency and thoroughness, we favor efficiency. And we prefer a practical approach to taxonomy too; adding a species on our site does not require valid peer-reviewed publication. And analysts can refer to "manuscript names" and communicate about yet-to-be described species. Our key tactic for "fixing" taxonomic issues is to have both site administrators and users illustrate problems or inconsistencies in published or unpublished taxa. Thus, our site promotes two-way communication with users to improve taxonomic consistency.



An early sketch or "wireframe" of a species page. These wireframes enabled the community of diatom researchers to provide feedback early in the design process, invaluable for subsequent development.

2. Structure, layout & interaction design

With a firm strategy and articulated approach in hand, the team next outlined how the site would be structured and arranged, with a continued emphasis on the user experience. This phase looked at the content and how it is best structured for quick comprehension, including what information is most important and needs to be highest in the visual hierarchy. The goal being to communicate science, not just organize information.

A key product of this organizing process was the creation of an extensive set of interactive "wireframes." These interlinked and sketched pages helped us design the navigation, user interface, and page layout. And equally important, the wireframes gave us an invaluable aid for gaining rapid and early feedback on our approach.

The initial wireframes were shown by Spaulding to about 50 participants at the North American Diatom Symposium in September 2007. She received considerable feedback. And at the same meeting, our collaborator Potapova (Academy of Natural Sciences Philadelphia), led a discussion with about 100 participants about the potential role of this web resource for the community. The response was quite positive. The wireframes were posted online at http://instaar.colorado.edu/diatoms which enabled us to gain additional feedback from diatom researchers not in attendance.

Based on such feedback, we decided to use a single mode for the user interface, not separate modes for "expert" and "non-expert" users. We also attempted to add a phylogenetic classification so that the site architecture and database would reflect diatom biology (evolutionary lineages). But after extensive discussions and review, we determined that such an approach is premature. More scientific progress is required before we can consistently apply phylogenetic classification. Therefore, we decided to structure the diatom genera based on "morphological groups."

The wireframes went through a series of iterative changes as we continued to get feedback from members of the diatom research community as well as perform informal usability testing. Several people suggested we make the interface more interactive, specifically to speed identification by showing supporting information and images. We then added information and images that would appear when rolling the mouse over key areas of the key pages or clicking on particular links and buttons.

3. Content creation & information flow

With a website structure in place, the team turned toward creating and managing content. Under the direction of Spaulding, team members began assembling and editing taxonomic information and imagery. Although most of the content came from existing sources, the effort was more time intensive than anticipated. For example, there were more than 75 separate pieces of information and images to gather, review, and track for each species alone. Moreover, more genera information was added to the project than originally planned to improve identification. So much so that Spaulding's efforts with helping assemble information on 110 genera meant she was unable to work on a large group of species. Nonetheless, our collaborator Potapova was able to create complete records for 25 species, more than sufficient for testing database structures and the function of the website. This change of plan did not alter the main project emphasis, which was to build a workable pilot site.

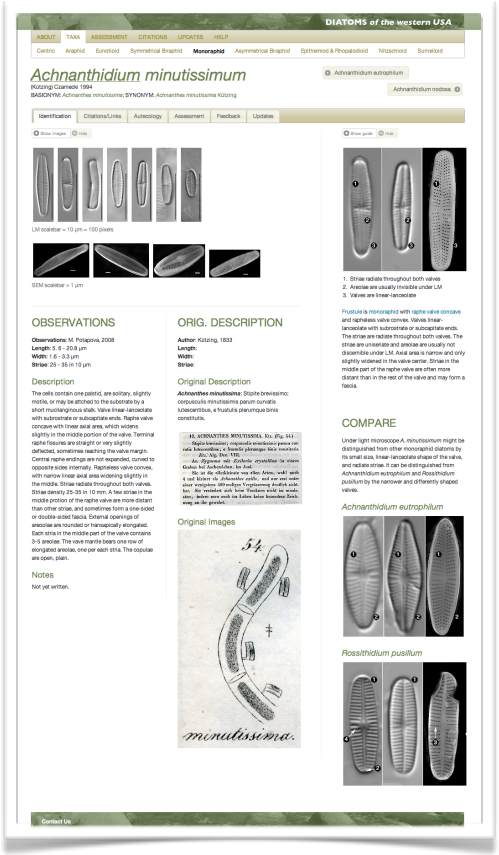
Besides assisting with taxonomic information, Kinsey also worked on obtaining permissions from publishers to include their original citation data and imagery in our website. Moreover, she worked on creating an extensive glossary too. Bell worked closely with Spaulding on genus level information and images.

Lubinski helped with this overall content creation process as well as providing expert advice on image processing, image metadata, and data entry forms/templates. Spaulding and Lubinski worked closely with Potapova to design and refine the templates for data entry. Together, we developed procedures for streamlined data entry via customized excel spreadsheets. We wanted to ensure that the time and effort to contribute data would not prevent diatom experts from participating.

4. Visual Design

Lubinski created the visual design for the website, with feedback from the rest of the team. A key design requirement was that the content must be prominent; it should not be overwhelmed by other graphic elements. In particular, a visual style was needed that worked well with grayscale imagery. Lubinski collected and reviewed a number of websites to help spark a solution. He drew much inspiration from the "A List Apart" website because of its unique combination of color highlights and grayscale imagery.

The key outcome of this step in the design process was a series of "mockups". These illustrations of key web pages were more refined than the sketchy wireframes but still faster to create and modify than coding actual web pages. We received additional feedback on these mockups, and also used them for additional informal usability testing.



A finished example of a species webpage, illustrating the visual design and the end result of some of the front-end development.

5. Front-end development: visuals & interactions

Lubinski coded all of the "front end" of the website, in other words, the portions actually seen by the user. He followed web standards, including those for Cascading Style Sheets (CSS), XHTML, Javascript, and RSS news feeds. All code was validated and tested in multiple browsers (Firefox, Safari, Internet Explorer). A semantic coding style will help with accessibility and search engine indexing. Coding efforts centered on creating templates that would be automatically "filled in" later with information from the database. Examples include one template each for species, genus, and morphological group.

Careful attention was given to how additional information would appear upon hovering the mouse over key images and links. This emphasis on interaction design will make the website considerably more functional and enjoyable to use. Informal usability testing was done to make sure the interactions were helpful, not confusing.

Spaulding demonstrated website front-end to participants in a special workshop at the 2009 North American Diatom Symposium. Approximately 40 diatomists attended her session to discuss the issues of taxonomic consistency and the database solution that we have constructed. Response of the scientific community was not only of overwhelming support, but urgency of need for this database solution.

6. Back-end development: engineering & database

Lubinski built all of the "back end" of the website, in other words, the code that controls what information is pulled from the database as well as the database itself.

Although the original intent was to modify previous PHP code and a MySQL database schema that Lubinski created for our Antarctic Diatoms website (http://huey.colorado.edu/diatoms), we deviated substantially from that plan. We realized that the long-term success of the western USA diatoms site meant that we had to make it simpler to update than the Antarctic site, which requires making changes to a fairly complicated relational database, FTP file uploads, and other somewhat technical procedures. Moreover we needed to add blog-like functionality to help promote communication with peers to help "fix" diatom identification problems. The easiest solution to all of those issues was to use a Content Management System (CMS). We thus re-aligned our back-end development plan to center on a CMS.

Lubinski chose a commercial CMS called ExpressionEngine, which runs on PHP and MySQL . ExpressionEngine was picked from the hundreds of CMS's on the market for its unique flexibility, strong user

community, great support, and quality of third-party add-ons. Lubinski had also used it previously to build a website for the CU-Boulder Environmental Studies Program (http://envs.colorado.edu). Before committing to ExpressionEngine, Lubinski verified that the performance of the software was in compliance with specifications.

The switch to ExpressionEngine for the western US diatoms project allowed Lubinski to build a simpler database schema than before, but with similar abilities to inter-relate information such as species, images, citations, genera, and morphological groups. The switch also enables our diatom experts to input their data directly into the system via easy-to-understand web-based forms, thereby saving time and reducing errors. These forms were created by Lubinski based on the custom excel "data templates" carefully designed by Spaulding, Lubinski, and Potapova earlier.

7. Entering content into the database

Although ongoing entry of information into the database/CMS will be done directly by content contributors themselves, the initial entry is being done by Lubinski. This one-time initial process is based on the existing excel data and also requires moving imagery into the ExpressionEngine file system so that relationships that inter-connect information and information are preserved.

The initial information was quality checked in three different ways before entry. First, Spaulding and Lubinski worked with the team during its early design phases to ensure that the excel-based "data templates" inherently reduced errors or made them easier to spot. For example, we used techniques such as fixed choices and limited field lengths. Second, Spaulding carefully reviewed the data for potential errors in entry, transcription, calculation, reduction, and transformation. The third method of quality control was done by Lubinski, who set up additional constraints for the online data entry such as required fields. Such constraints promote data integrity and completeness.

New information entered directly into the online database will undergo a formal approval process. Diatom experts will submit information themselves but that data will not be displayed to the public until approved by a reviewer (initially Spaulding). The reviewer will make corrections and sometimes ask the expert contributor to reconsider information or make changes. After all is resolved, the reviewer will witch the status of the data from "pending" to "approved."

A web-based form that diatom specialists would use to enter species data via their web browser. There are eight tabs for this form, the second of which is open in this example. Several rarely used fields are deliberately not shown until the user clicks the disclosure triangle. This approach helps keep the interface from getting too cluttered. Note: colors and layout of this form are likely to change before site launch.

8. Testing & training

Lubinski tested the both the front- and back-end portions of the site throughout development. An additional focussed testing effort at the end of development ensures that all is functioning properly, and that there are no broken links, etc. Subsequent training sessions with content contributors not only provide invaluable "in-person" instructions but often reveal minor issues that can be fixed before launch. For example, this is a great opportunity to reword instructions to eliminate any confusion during data entry.

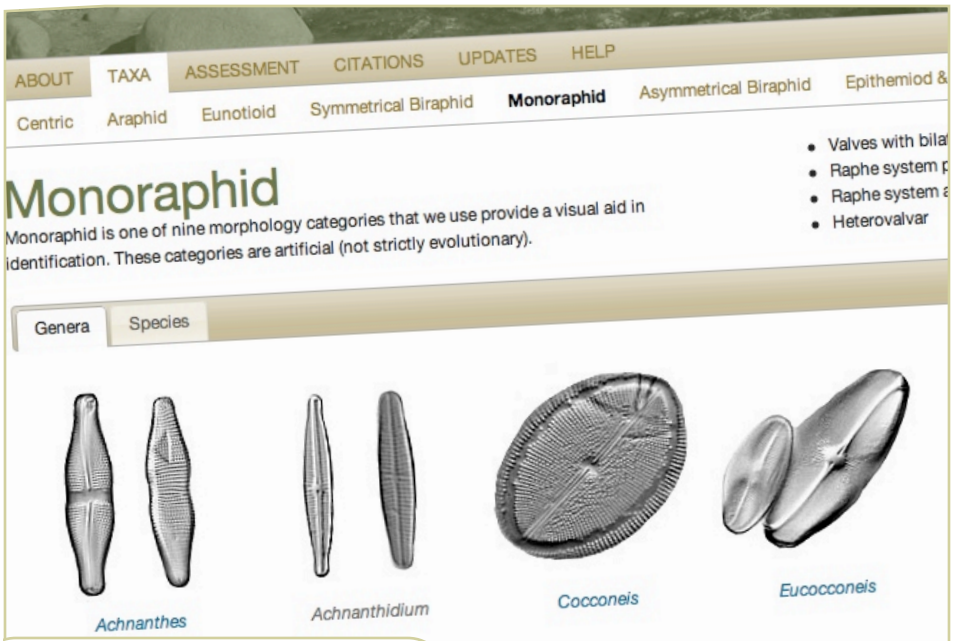
9. Launching & hosting

Lubinski has set up the site on a shared INSTAAR computer system at the University of Colorado. The system includes separate web and database servers to maximize performance. Lubinski will coordinate launch efforts with the university IT team and will help with announcements, particularly to ensure that search engines, like Google, will visit the new site for indexing as soon as possible. Once the site is indexed, Lubinski will set up Google-powered site search. We anticipate continuing to host the site on the INSTAAR servers during the next phase of funding (see last page of report), although the web address URL may change.

10. Reviewing & followup

After launch, the site will be reviewed by the five members of our project Data Review Board, EPA representatives for the Region 8 states, and stakeholders from western states (see remaining tasks on last page of this report). The reviews will be used by Lubinski to make minor site modifications. Larger changes will be considered for future releases.

After the site has been in use for some months by the diatom community, Lubinski and Spaulding will have a followup meeting to assess failures, successes, lessons learned, and remaining questions. We will look at web traffic patterns too. This session will help guide future site improvements. Another valuable improvement method is Google's free A/B testing service. With it, we can deliver one version of a page to half of the site's visitors and a slightly different version to the other half. We can then determine which page is easier to use by tracking patterns of use, such as time on the page, which links were clicked, and interviews with users.



PROJECT STATUS

Mostly complete, but several key tasks remain. All will be finished at no additional cost. The project deliverable is a website, which will be fully reviewed and operational by february 15th 2010 at westerndiatoms.colorado.edu

Finish development by January 1, 2010

Lubinski will complete the database, including final entry of 110 genus records and 25 species records. And he will document the web resource. Database structure and documentation will be available directly from the website under a clearly marked section labeled "documentation" or equivalent.

Board & EPA review by january 15, 2010

Review of database by Data Review Board (M. Edlund, P. Kociolek, K. Manoylov, K. Hermann, B. Beyea). All reviewers will be asked to provide feedback on strengths and weaknesses of the project in addressing the project objectives. Reviewers will be asked to recommend or decline further funding for the project. EPA representatives of Region 8 states will be asked to test the database system for ease of use, information content, and ability to address taxonomic consistency. After the reviews, Lubinski will make site modifications, and take notes on larger-scale changes to be made in the future.

State review by February 1, 2010

Spaulding, with the support of EPA Region 8, will request Region 8 states to critically evaluate the functioning of the web resource for ease of use, information content, and ability to address taxonomic consistency. Region 8 states will be asked to give final approval that Spaulding has completed the objectives of the project. After the reviews, Lubinski will make site modifications, and take notes on larger-scale changes to be made in the future.

System operational by February 15, 2010

The system will begin accepting contributions for additional species submission, monitored by Spaulding. We have received commitment of additional funding from USGS Water Resources Division (WRD) and EPA Office of Water (OW) for further expansion of the database with additional records and links to other electronic floras (see text in the sidebar to the right).

FUTURE PLANS

Transition the website from regional to national in scope

Expansion

We had originally intended this site to be a western site. However, many of our colleagues and peers would like the site to become national in scope. Therefore, Spaulding has sought additional funding to support a greater number of expert contributors and taxon pages, changes in database structure to reflect national geographic coverage, and streamlining of data submission.

Funding

Funding is expected from several sources in 2010. USGS WRD is budgeting for database expansion to a national level and population of species taxon pages ($100K). EPA OW is also expected to contribute $100K to fund expert contributions of taxon pages. The Academy of Natural Sciences (ANSP) cooperative agreement with USGS NAWQA is committing resources for Potopova to develop 90 additional taxon pages. USGS NAWQA has also committed 6-9 months of salary for Spaulding to coordinate the database effort.

CLOSING SUMMARY

This project is one of the first of its kind for diatom taxonomy. We believe that we have produced a database framework and design that can be expanded in the future. Documentation of species diversity is of great concern and there are numerous efforts and new technologies that are emerging. We intend to stay informed about changes and incorporate new database and computing tools in our next efforts. In the end, we hope we have created a tool that accommodates corrections to content, expands with new knowledge, and grows as a resource.